BEGINNING OF TRANSMISSION

State: published
Pairing: complete (DOI, GitHub, SHA-256, Zenodo URL)

Release metadata

Field	Value
Title	On-Policy Distillation as Active Inference in Finite Variational Models
Version	1.0.2
Concept DOI	10.5281/zenodo.20747834
Version DOI	10.5281/zenodo.20749817
GitHub	https://github.com/ActiveInferenceInstitute/on_policy_distillation/releases/tag/v1.0.2
Zenodo	https://zenodo.org/records/20747834
SHA-256	`8d70985fca938772…`
SHA-512	pending

How to verify

- Scan **Integrity** QR and compare the embedded SHA-256 prefix to the table above.
- Scan **Zenodo** / **GitHub** QR codes and confirm they resolve to this release pairing.
- Full hashes and structured fields: `../data/transmission_manifest.json`.

Figure 1: Integrity QR strip

Structured manifest: `../data/transmission_manifest.json`

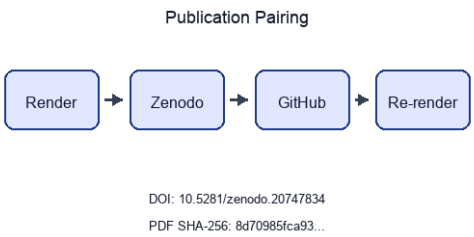


Figure 2: Publication pairing flow

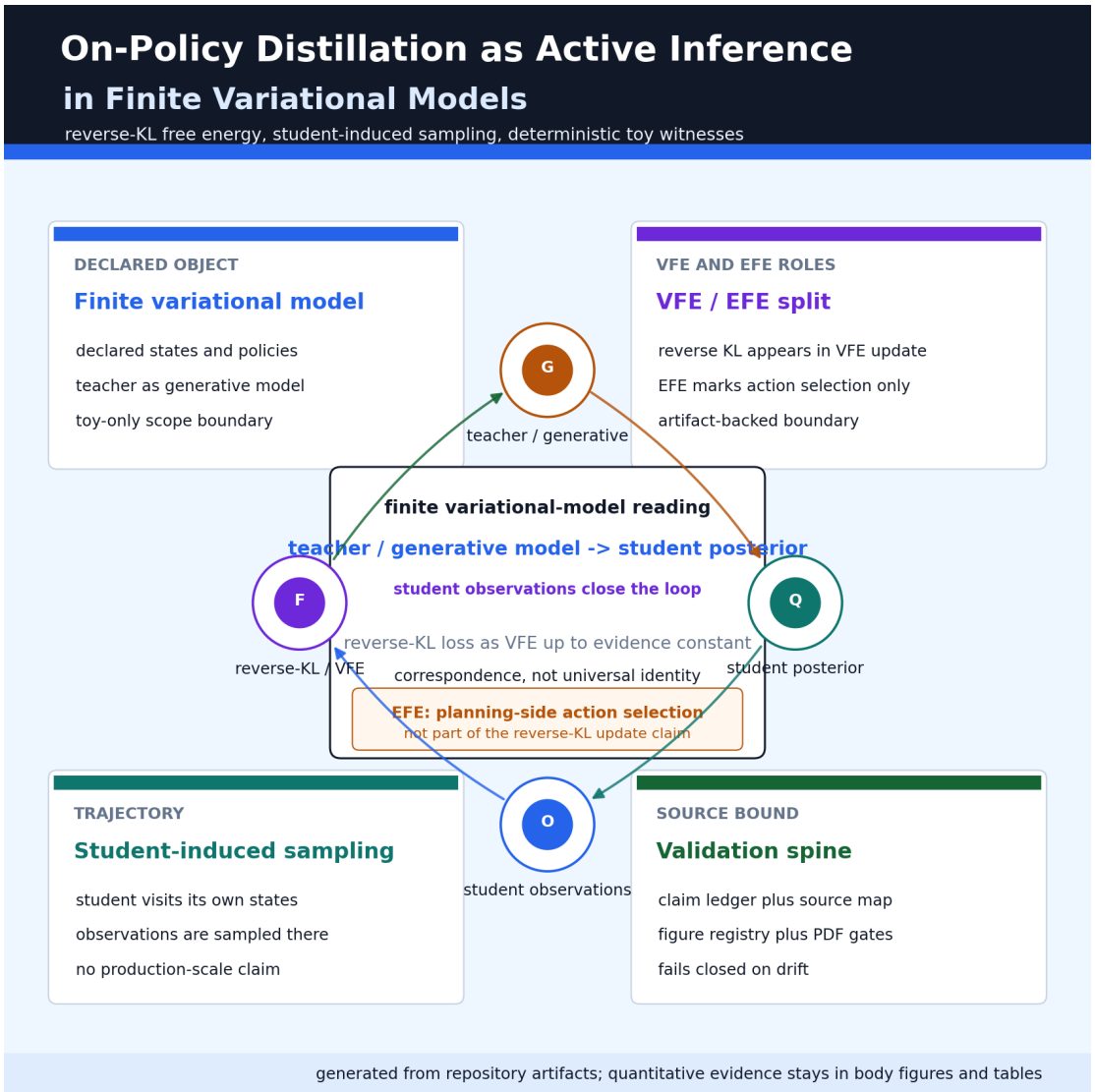
Stego: off | overlays text | barcodes on | XMP on | manifest on → `./secure_run.sh`

On-Policy Distillation as Active Inference in Finite Variational Models

Reverse-KL free energy, student-induced sampling, and deterministic toy witnesses

Daniel Ari Friedman
Active Inference Institute
daniel@activeinference.institute
ORCID: 0000-0001-6232-9096
DOI: 10.5281/zenodo.20747834

June 2026



Contents

1 Sheaf Track Coverage **3**
- 1.1 Introduction . . . 3
- 1.2 Methods . . . 3
- 1.3 Results . . . 3
- 1.4 Discussion . . . 3
- 1.5 Appendix . . . 3

2 Abstract **4**

Introduction **5**

3 Motivation and scope **5**
- 3.1 Scientific scope . . . 5
- 3.2 Manuscript structure . . . 6

4 Contributions **7**
- 4.1 Scientific contributions . . . 7
  - 4.1.1 Ontology bindings . . . 9

Methods **12**

5 Teacher and student coupling: the analytical model **12**
  - 5.0.1 Ontology bindings . . . 14

6 On-policy student: pymdp sophisticated inference **15**
  - 6.0.1 Ontology bindings . . . 16

7 Machine-checked correspondence (Lean) **18**
  - 7.0.1 Proof extraction track . . . 20

Results **21**

8 Teacher and student mutual information **21**

9 Free-energy decomposition **22**
  - 9.0.1 Energy decompositions: VFE and EFE . . . 22

10 On-policy student rollout (T-maze) **27**

Discussion **33**

11 Limitations and outlook **33**
- 11.1 What this supports . . . 33
- 11.2 Limitations . . . 33
- 11.3 Threats to validity . . . 34
- 11.4 Empirical evidence (literature-reported) . . . 34
- 11.5 Audit, evidence, and open problems . . . 34
- 11.6 Toward LLM and world-model training runs . . . 35
  - 11.6.1 Ontology bindings . . . 36
  - 11.6.2 Release notes evidence track . . . 37

12 Conclusion **38**

Appendix **39**

13 Supplementary material: full coverage and concordance **39**
  - 13.0.1 Supplemental table: energy decomposition . . . 39
  - 13.0.2 Supplemental table: empirical OPD-vs-RL benchmark (literature-reported) . . . 39
  - 13.0.3 Appendix track: artifact diffoscope . . . 40
  - 13.0.4 Appendix track: artifact license . . . 40
  - 13.0.5 Appendix track: state-space catalog . . . 41
  - 13.0.6 Appendix track: causal ablation . . . 43

14 Supplementary material: reproducibility methodology **44**
14.1 Compose contract . . . 44
14.2 Coverage and figures . . . 44
14.3 Compose commands . . . 44
14.4 Law verification . . . 44
14.4.1 Base poset and presheaf . . . 45
14.4.2 Verified sheaf laws . . . 45
14.4.3 Scope (what is and is not claimed) . . . 45
14.4.4 Artifact diffoscope track . . . 49
14.4.5 Artifact license track . . . 49
14.5 Sheaf fragment track registry . . . 50
14.6 IMRAD binding matrix . . . 51
14.7 Section-track status . . . 53
14.8 Track status . . . 54
14.9 Render and logging summary . . . 54
14.10 Evidence crosswalk . . . 55
14.11 Artifact producer graph . . . 55
14.12 Semantic gluing restrictions . . . 59
14.13 Track improvement scope . . . 59

15 Supplementary material: validation invariants and statistics **62**
15.0.1 Appendix track: proof extraction . . . 63
15.0.2 State-space catalog track . . . 63
15.0.3 Causal ablation track . . . 63
15.0.4 Ontology bindings . . . 63
15.0.5 Appendix track: release notes evidence . . . 64

16 References **65**

1 Sheaf Track Coverage

This page summarizes which **sheaf fragment tracks** are bound for each IMRAD row in `manuscript/sheaf/manifest.yaml`. The matrix is regenerated at compose time.

Totals: 95 present / 95 bound / 0 missing (gray).

Color	Meaning
Black	Track **present** (bound and fragment exists)
White	**Absent** (not bound for this row)
Gray	**Missing** (bound but fragment file absent)

1.1 Introduction

- **Introduction** *(group)*
- **Motivation and scope**
- **Contributions**

1.2 Methods

- **Methods** *(group)*
- **Teacher and student coupling: the analytical model**
- **On-policy student: pymdp sophisticated inference**
- **Machine-checked correspondence (Lean)**

1.3 Results

- **Results** *(group)*
- **Teacher and student mutual information**
- **Free-energy decomposition**
- **On-policy student rollout (T-maze)**

1.4 Discussion

- **Discussion** *(group)*
- **Limitations and outlook**

1.5 Appendix

- **Appendix** *(group)*
- **Supplementary material: full coverage and concordance**
- **Supplementary material: reproducibility methodology**
- **Supplementary material: validation invariants and statistics**

Appendix row `18_supplement_full_coverage.md` binds 22 fragment track types as a composability proof (registry defines 33 types; generated `layers` included).

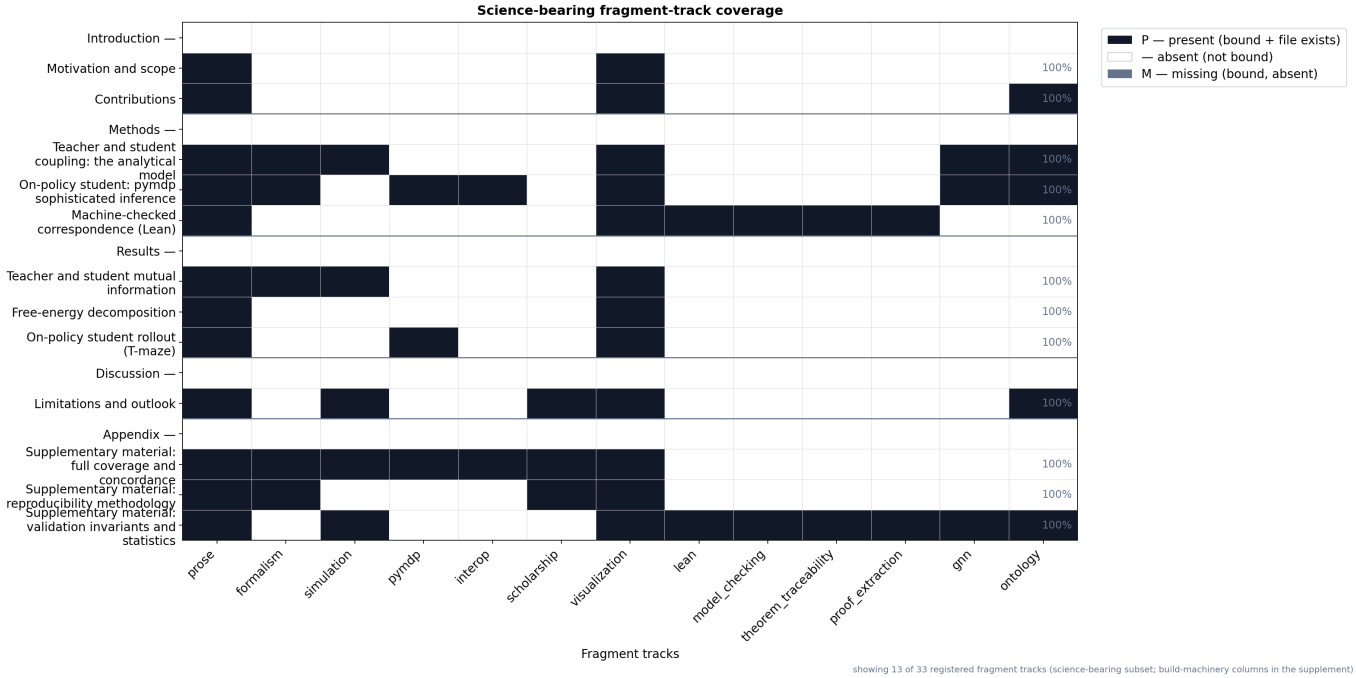


Figure 3: Sheaf track coverage matrix mapping 17 IMRAD manuscript rows against 33 composable fragment-track columns: black = present (P), white = absent (—), gray = bound-but-missing (M), with counts 95 present / 95 bound / 0 missing. The figure shows the science-bearing track subset; build-machinery columns are omitted and the full matrix appears in the supplement. The matrix is the gluing record showing which evidence fragment is locally attached to each manuscript section. Reading it as a sheaf condition, a consistent (fully present, zero-missing) column set is what licenses gluing the local toy results, Lean witnesses, and pymdp rollouts into one globally coherent active-inference argument about on-policy distillation.

2 Abstract

This paper formulates on-policy distillation as active inference in finite variational models, with exact claims only for declared objects and interpretive claims explicitly bounded outside them. In the construction, the intractable teacher policy plays the role of the generative model $p(o, s)$, the tractable student policy is the approximate posterior $q(s)$, and the per-token reverse-KL distillation loss is variational free energy up to the evidence constant, $F = D_{\mathrm{KL}}(q \,\|\, p(s \mid o)) - \log p(o)$, whose KL target is the teacher-induced posterior $p(s \mid o) \propto p(o, s)$ [Friston et al., 2009, Friston, 2010, Parr et al., 2022, Levine, 2018].

The title's "as" is therefore a scoped mathematical correspondence rather than the slogan `OPD = Active Inference`. Variational free energy names the realized-rollout distillation loss; expected free energy remains the planning-side objective by which the pymdp agent selects actions [Millidge et al., 2021b, Friston et al., 2021a, Da Costa et al., 2020]. On-policy student rollouts generate the observations on which the posterior is scored, connecting the construction to induced-distribution mismatch in imitation learning and exposure-bias analyses while preserving their different objectives, empirical regimes, and contested severity [Ross et al., 2011, Bengio et al., 2015, Huszár, 2015, Ranzato et al., 2016, He et al., 2021]. Privileged traces and feedback play the role that train-time-only information plays in the LUPI/distillation lineage [Vapnik and Vashist, 2009, Lopez-Paz et al., 2016, Sharoni and Sabato, 2023].

Four deterministic witnesses instantiate the correspondence. A Bernoulli-Ising oracle couples a teacher's privileged variable to the answer through λ, making $I(\lambda)$ the teacher-student mutual information and the finite free-energy gap the toy distillation objective; the closed-form and independently recomputed mutual-information sweeps agree to machine precision (RMSE 2.1e-16 nats). A pymdp T-maze rollout supplies the on-policy student that samples its own observations under a privileged cue [Friston et al., 2021a, Millidge et al., 2021b, van Oostrum et al., 2024]. A two-agent classroom pits a privileged teacher (cue validity 0.98) against an on-policy student (cue validity 0.5), measuring teacher belief entropy 0.247 nats versus student 0.347 nats and a mean reverse-KL distillation signal of 6.28 nats. A four-state/two-action sequential-shift witness shows teacher-forced train loss 0.333 nats underestimating student-induced test loss 0.409 nats, with deterministic on-policy correction reducing it to 0.096 nats.

These are toy, generated findings, not production-LLM measurements. Recent privileged-context, context-distillation, adaptive-teacher, freshness-aware OPD, RLHF/instruction-tuning, self-generated reasoning, Qwen OPD-vs-RL, and Thinking Machines replication reports remain external context rather than reproduced results [Zhao et al., 2026, Liu et al., 2026e, Penaloza et al., 2026a, Snell et al., 2022, Ye et al., 2026, Lazaridis et al., 2026, Han et al., 2026, Chen et al., 2026, Ouyang et al., 2022, Zelikman et al., 2022, Qwen Team, 2025, Lu and Thinking Machines Lab, 2025]. The supplemental sheaf/provenance layer keeps that boundary operational: every reported number is hydrated from a generated artifact, every figure is source-bound, and 16 / 16 invariant checks pass before rendering.

Introduction

3 Motivation and scope

3.1 Scientific scope

A student policy trained only on a teacher's own outputs learns under a distribution it will not induce at inference time. That is the sequential-prediction failure mode behind behavioral cloning [Pomerleau, 1989], efficient imitation-learning reductions [Ross and Bagnell, 2010], DAgger [Ross et al., 2011], differentiable interactive imitation [Sun et al., 2017], scheduled sampling [Bengio et al., 2015], sequence-level training objectives [Ranzato et al., 2016], and language-generation exposure-bias analyses [Arora et al., 2022, Rohatgi et al., 2025]: each generated token or action changes the next state distribution, so errors can compound precisely where the learner has not been trained.

Two caveats keep this lineage honest. Scheduled sampling is a historically important response to teacher-forcing mismatch whose objective has itself been criticized as statistically inconsistent, so we cite it as part of the repair lineage rather than as a settled solution [Huszár, 2015]. The empirical severity of exposure bias is also task-dependent — autoregressive models can exhibit meaningful self-recovery — so our use of the term is motivational rather than universal [He et al., 2021]. Privileged-information theory adds a second caution: train-only information can help, but unrestricted privileged capacity need not improve worst-case guarantees, so this manuscript treats privilege as a finite artifact field rather than as a free generalization theorem [Vapnik and Vashist, 2009, Sharoni and Sabato, 2023].

Passive, off-policy supervised distillation descends from model compression and classical KD [Buciluǎ et al., 2006, Hinton et al., 2015, Stanton et al., 2021], sequence-level KD [Kim and Rush, 2016], and policy-distillation work in RL [Rusu et al., 2016, Czarnecki et al., 2019], but it inherits that mismatch when it minimises a teacher-data objective on trajectories drawn from the teacher rather than the student. On-policy distillation targets the mismatch by scoring the student's *own* samples under the teacher and minimising a reverse, skew, entropy-aware, contrastive, or hybrid KL on the induced student distribution [Gu et al., 2024, Agarwal et al., 2024, Ko et al., 2024, 2025, Wu et al., 2024, Jin et al., 2026, Zhu et al., 2026b]. Recent OPD studies sharpen the boundary of that intervention: teacher choice, loss formulation, teacher entropy, privileged-information type, context-window transfer, black-box access, trust regions, adaptive targets, and long-horizon reward coupling can determine whether the on-policy signal stabilises learning or introduces new failure modes [Pozzi et al., 2025, Oh et al., 2026, Xing et al., 2026, Jang et al., 2026].

A closely parallel correction — related in mechanism though differing in objective and empirical regime — is the defining move of active inference as a process theory: an agent acts to generate observations under which its approximate posterior is refined, and epistemic policies deliberately sample cues that reduce uncertainty [Friston et al., 2006, 2009, Friston, 2010, 2013, Friston et al., 2017a,b, Sajid et al., 2021a, Friston et al., 2021b, Aguilera et al., 2022, Parr et al., 2022, Da Costa et al., 2020, Tschantz et al., 2020a, van Oostrum et al., 2024].

The paper's thesis is therefore a scoped reading of declared formal objects, not a loose metaphor and not a process-level identity for every OPD system. In the finite models studied here, the teacher policy plays the role of the intractable generative model $p(o, s)$, the student policy is the tractable approximate posterior $q(s)$, the variational free energy $F = D_{KL}(q \,\|\, p(s \mid o)) - \log p(o)$ (KL target the exact posterior $p(s \mid o) \propto p(o, s)$) is the per-token reverse-KL distillation loss, and on-policy student rollouts are the active sampling that lets the posterior generate its own observations.

Privileged information available in training but not at inference - a hint, verified trace, feedback channel, long context, visual cue, or layer-internal predictive signal - is useful only when it transfers into the student's deployment variables rather than becoming a shortcut the student cannot use later. In this manuscript that privileged access is modeled as a teacher-side conditioning variable across a blanket-like conditional-independence partition — a constrained probabilistic reading only, since blanket-based inferential interpretations are technically delicate and contested in general, and no physical or biological boundary claim is made [Kirchhoff et al., 2018, Biehl et al., 2021, Aguilera et al., 2022]. That links learning under privileged information [Vapnik and Vashist, 2009, Sharoni and Sabato, 2023], distillation-as-privileged-information [Lopez-Paz et al., 2016], context distillation [Snell et al., 2022, Ye et al., 2026], privileged/contextual OPD [Zhao et al., 2026, Liu et al., 2026e, Penaloza et al., 2026a, Lazaridis et al., 2026, Liu et al., 2026a], and internal on-policy self-distillation [Liu et al., 2026c]. Predictive-coding language is used in the same limited way: the teacher supplies a top-down target distribution and the student updates from the residual on its own generated trajectory, echoing hierarchical prediction-error correction without claiming cortical implementation [Rao and Ballard, 1999].

Self-Distillation Fine-Tuning also names two distinct lines that this manuscript keeps separate: Yang et al.'s ACL SDFT uses model-generated distilled data to bridge a fine-tuning distribution gap [Yang et al., 2024], whereas Shenfeld et al.'s SDFT uses a demonstration-conditioned model as its own teacher for continual learning [Shenfeld et al., 2026].

The model surface is deliberately small and named. **Bernoulli-Ising** is the analytical teacher-student coupling oracle: it supplies the closed-form mutual-information and free-energy calculations. **pymdp T-maze** is the on-policy active-inference rollout: it supplies the agent that samples its own observations under sophisticated inference. **Classroom** is the two-agent teacher/student distillation signal: it compares a privileged teacher against an on-policy student on the same toy task. **Sequential-shift** is the finite distribution-shift witness: it compares teacher-forced train visitation with student-induced test visitation and a deterministic on-policy correction. **Graph-world** is a finite topology stress-test and Lean/model-checking extension: it checks small reachability and artifact-consistency boundaries rather than serving as the main empirical environment. No gridworld result is reported or claimed. The conceptual lineage is the free-energy and active-inference literature [Friston et al., 2006, 2009, Friston, 2010, Sajid et al., 2021a, Friston et al., 2021b, Aguilera et al., 2022, Parr et al., 2022], but the scientific claims stay within these models and their generated artifacts — they are not empirical statements about biological agents or production language models.

3.2 Manuscript structure

Three **scientific tracks** — analytical, pymdp, and the formal/publication track (Lean, sheaf composition, provenance; the third lane of fig. 7) — map onto 33 **composable fragment types** and 30 required pipeline tracks (fig. 7); each instantiates one face of the teacher–student correspondence rather than standing as an isolated exhibit. The Bernoulli–Ising analytical oracle is a minimal model of teacher–student coupling: a coupling λ ties the teacher's privileged variable to the answer, the mutual information $I(\lambda)$ is the teacher–student mutual information, and the mean-field free-energy gap is the distillation objective the independent student must close.

The pymdp T-maze rollout is the on-policy student itself: an agent that generates its own observations and acts to minimise expected free energy, where the cue is the privileged information and its validity sets how privileged it is. A two-agent classroom simulation closes the loop, running a privileged teacher against an on-policy student on the same task; its belief-entropy gap and reverse-KL distillation signal are measured in sec. 10. A sequential-shift witness then checks the review-requested train/test mismatch in a four-state, two-action finite system, with a correction-dose sensitivity sweep to keep the result from depending on one hand-picked correction level.

These executable demonstrations are placed beside, not substituted for, external OPD evidence and surveys [Agarwal et al., 2024, Lu and Thinking Machines Lab, 2025, Liu, 2026, Song and Zheng, 2026, Ko et al., 2024, Jin et al., 2026, Zhu et al., 2026b,a, Ramos et al., 2026]; every quantitative claim below remains a claim about this project's generated artifacts. sec. 1 summarizes which fragment tracks bind to each manifest row. The standalone reproducibility supplement (sec. 14) documents the compose pipeline, coverage semantics (eq. 6), and strict validation gates.

The pymdp track follows pymdp's sophisticated-inference TMaze validation profile [Heins et al., 2022] with the full TMaze environment, SI search horizon 5, and Agent `policy_len = 1`; sophisticated inference — beliefs about beliefs in a deep temporal generative model — is the toy formal counterpart of a teacher conditioned on the student's own verified traces or internal predictive states [Friston et al., 2018, Shenfeld et al., 2026, Hübotter et al., 2026].

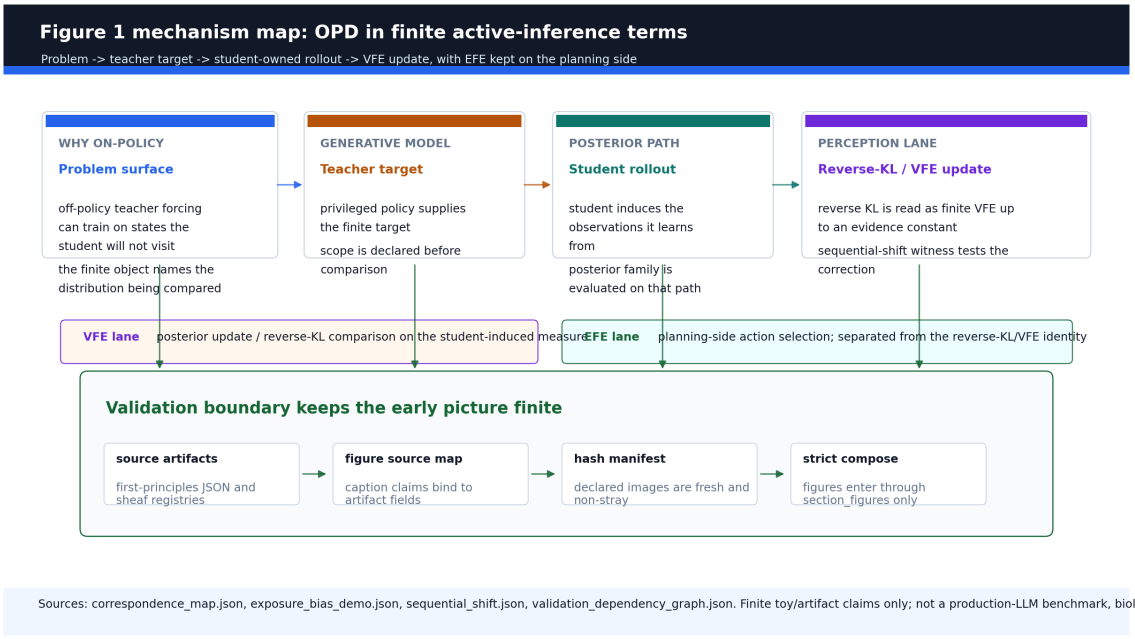


Figure 4: Source-bound Figure 1 mechanism map for the manuscript's opening sections. The figure is a schematic finite reader map, not a metric dashboard: it shows the path from problem surface to teacher target, student-owned rollout, and reverse-KL/VFE update while keeping EFE on the planning-side action-selection lane. The bottom boundary and validation dependency graph state that the figure supports finite toy/artifact claims only; figure provenance, hash manifests, and strict compose gates are part of the claim, not decoration. Sources: `output/data/firstprinciples/correspondence_map.json`, `output/data/firstprinciples/exposure_bias_demo.json`, `output/data/firstprinciples/sequential_shift.json`, and `output/data/validation_dependency_graph.json`.

4 Contributions

4.1 Scientific contributions

We argue that on-policy distillation (OPD) admits an active-inference reading at the level of the finite variational objects studied here: the intractable teacher policy $\pi_T(y \mid x, I)$ plays the role of the generative model $p(o, s)$, the tractable student family $\pi_S(y \mid x)$ plays the role of the approximate posterior $q(s)$, and the per-token reverse-KL distillation loss is the variational free energy $F = D_{\mathrm{KL}}(q \| p(s \mid o)) - \log p(o)$, the KL target being the exact posterior $p(s \mid o) \propto p(o, s)$ the teacher induces [Kullback and Leibler, 1951, Jordan et al., 1999, Blei et al., 2017, Friston et al., 2006, 2009, Friston, 2010, Friston et al., 2017a, Parr et al., 2022]. This paper substantiates that scoped correspondence with an audited map, executable minimal models, and a source-bound manuscript pipeline. We make five contributions.

Because the correspondence relabels objects across two vocabularies, we fix the type translation once, before any derivation. The mapping below is applied only after this translation; the teacher/student reading is never assumed implicitly.

Symbol	Active-inference role	On-policy-distillation role
x	conditioning context	prompt
y	— (output index)	generated token / sequence
I	privileged information	teacher-only signal (verified trace / cue)
o	observation	realized token / outcome
s	hidden state	latent the student must infer
$p(o, s)$	generative model	teacher policy $\pi_T(y \mid x, I)$
$q(s)$	approximate posterior	student policy $\pi_S(y \mid x)$
$p(s \mid o)$	exact posterior (KL target)	teacher-induced target
F	variational free energy	per-token reverse-KL loss
G	expected free energy	data-collection / planning side, not the realized-rollout loss
λ	teacher–student coupling	strength of teacher signal
β	precision / temperature	distillation temperature

We also fix, up front, exactly what kind of claim each result is, so that a reader can separate what is proved from what is illustrated from what is borrowed as context. Every assertion in the paper falls into one of four tiers, formalized in the scoped Proposition (sec. 5):

Tier	Claim kind	Status	Where
1	**Algebraic identity** — reverse-KL distillation loss equals variational free energy up to the evidence constant; the mutual-information / conditional-entropy complement	proved in closed form (two-route verified)	Proposition (i)–(ii)
2	**Numerical witness** — reverse-KL and free-energy descent reach the same posterior; the pymdp T-maze, classroom, and sequential-shift toys	measured on deterministic toy artifacts	Proposition (iii), sec. 10
3	**Interpretive analogy** — on-policy rollouts as active sampling, differential cue reliability as privileged information, expected free energy as planning, the Markov-blanket reading	a correspondence built on Tiers 1–2, not a further theorem	Proposition (iv), sec. 11
4	**External context** — literature-reported OPD-vs-RL empirics (Qwen3, Thinking Machines)	not measured here; cited only as neighbouring context	sec. 11

1. **Audited correspondence map** (sec. 5): a checked, component-by-component identification of the active-inference machinery with the OPD machinery — generative model to teacher, posterior to student, free energy to reverse-KL loss, active sampling to on-policy student rollouts (the posterior generating its own observations), epistemic value to teacher signal on novel student states,

pragmatic value to the reward-tilted target $\exp(R/\beta)$, the Markov blanket to teacher/student context asymmetry, predictive coding to top-down teacher target plus bottom-up residual, privileged sensory access to the privileged information I, and sophisticated inference to a teacher conditioned on the student's own verified traces [Friston et al., 2017b, Da Costa et al., 2020, Sajid et al., 2021a, Friston et al., 2018, 2021a, Kirchhoff et al., 2018, Rao and Ballard, 1999, Vapnik and Vashist, 2009, Lopez-Paz et al., 2016, Zhao et al., 2026, Liu et al., 2026e]. The correspondence is exact for the explicitly constructed finite toy objects studied here — a claim we pin down as a proposition with stated assumptions in sec. 5, separating what is proved in closed form, what is demonstrated numerically, and what remains an interpretive reading — and we keep all claims scoped to these minimal models and artifacts. The full dictionary — all 26 machine-validated rows — is rendered as fig. 6, so the thesis can be read as a single picture before any derivation.

2. **Shared divergence geometry** (sec. 5): closed-form mutual information $I(\lambda)$ and a free-energy decomposition on a symmetric Bernoulli-Ising toy, with an independent exact-recomputation cross-check (sec. 8, sec. 9). Here λ couples the teacher's privileged variable to the answer, $I(\lambda)$ is the teacher-student mutual information, and the finite free energy is the distillation objective for this toy. The entangled posterior versus mean-field comparison instantiates the divergence-direction choice that organises the OPD landscape [Liu, 2026]: the reverse-KL side concentrates on target-supported mass in this finite example (the MiniLLM/GKD lineage [Gu et al., 2024, Agarwal et al., 2024]), the forward-KL side covers teacher mass (the SFT and classical knowledge-distillation limit [Hinton et al., 2015], with fine-tuning distribution-gap variants kept as context [Yang et al., 2024]), skew/adaptive KL methods occupy middle regimes [Ko et al., 2024, Jin et al., 2026, Zhu et al., 2026b], and alpha/f-divergence or KL-geometry work warns against treating that toy contrast as a universal LLM law [Hernández-Lobato et al., 2016, Ke et al., 2019, Wu et al., 2024]. Student-induced rollouts address the training/inference mismatch identified in sequential prediction and LLM exposure-bias work [Ross et al., 2011, Bengio et al., 2015, Arora et al., 2022, Pozzi et al., 2025].

3. **Reward-tilted-target unification** (sec. 5): we show that control-estimation duality [Todorov, 2008], trajectory inference [Toussaint, 2009], maximum-entropy IRL [Ziebart et al., 2008], variational intrinsic control [Fellows et al., 2019], RL-as-inference [Levine, 2018, Abdolmaleki et al., 2018], control-as-inference [Millidge et al., 2020a], maximum-entropy RL [Haarnoja et al., 2018], RLHF and DPO-style KL-constrained preference objectives [Ouyang et al., 2022, Ziegler et al., 2019, Rafailov et al., 2023], active inference, and on-policy distillation can all be written against related KL-regularized, reward-tilted targets of the form $\pi_{\text{ref}} \exp(R/\beta)$ — with pragmatic value entering as the reward tilt and the distillation temperature β playing the role of the precision γ — and are best treated as a structured family that differs in target construction, priors, and regularizers rather than as a single objective [Friston et al., 2017a, Millidge et al., 2021b, Penaloza et al., 2026a,b]. This places preference-feedback variants [Hübotter et al., 2026] and the privileged-context self-distillation objective $\mathcal{L} = \mathcal{L}_{\text{clip}} + \beta\, D_{\text{KL}}(\pi(\cdot \mid x) \| \pi(\cdot \mid c, x)) + \alpha\, \text{KL}_{\text{ref}}$ [Liu et al., 2026e,d] inside one variational frame.

4. **Two-agent pymdp classroom plus sequential-shift witness** (sec. 6): a deterministic pymdp full TMaze rollout under sophisticated inference as the canonical on-policy student - an agent that generates its own observations and acts to minimise expected free energy - with logged q_π rows, action marginals, modality observations, matrix/value audit, SI tree diagnostics, and merged invariant gates (sec. 10, supplement sec. 15). Privilege is operationalized here as *differential cue reliability* across agents — the cue observation plays the role of the privileged information I, and `cue_validity` sets how reliable each agent's access to it is — a structured-partial-observation analogue of LUPI-style train-only information rather than a literal variable removed at deployment [Vapnik and Vashist, 2009, Lopez-Paz et al., 2016, Sharoni and Sabato, 2023, van Oostrum et al., 2024]. We use the blanket-like conditional-independence partition only as a constrained probabilistic reading of teacher/student context asymmetry in these toys; blanket-based inferential interpretations are technically delicate and contested in general, and no broader physical or biological claim is intended [Friston, 2013, Kirchhoff et al., 2018, Biehl et al., 2021]. A two-agent "classroom" simulation pits a privileged teacher (`cue_validity` 0.98) against an on-policy student (`cue_validity` 0.5), measuring teacher belief entropy 0.247 nats versus the student's 0.347 nats - an entropy gap induced by the toy's privileged cue-validity asymmetry and relevant to teacher-entropy OPD - at a mean reverse-KL distillation signal of 6.28 nats [Tschantz et al., 2020a, Jin et al., 2026, Han et al., 2026, Chen et al., 2026]. The same results section includes `firstprinciples.sequential_shift.v1`, a deterministic four-state/two-action witness showing teacher-forced training underestimates student-induced test loss and an on-policy correction reduces it, plus `firstprinciples.sequential_shift_sensitivity.v1`, a finite correction-dose sweep that checks this improvement across policy mixtures; these are toy train/test shift checks, not production OPD benchmarks [Shimodaira, 2000, Sharoni and Sabato, 2023, Zelikman et al., 2022]. This is a local executable demonstration of the information the student's free energy must close, not a reproduction of production OPD efficiency, adaptive-teacher, or freshness-aware asynchronous OPD claims [Qwen Team, 2025, Lu and Thinking Machines Lab, 2025].

5. **Sheaf-indexed composition** (supplement sec. 14): 33 optional fragment types bind to 17 manifest rows under eq. 6, with a 22-track appendix composability proof (sec. 13), so the manuscript that states the correspondence is itself a gate-checked, sheaf-composed artifact whose claims are mechanically traceable. This is an applied composition-and-contract use of sheaf language [Curry, 2014, Speranzon et al., 2018, Robinson, 2014], while the active-inference diagrams and Generalized Notation Notation (GNN — the graphical model-specification language, not graph neural networks) and ontology round-trips are situated against graphical active-inference specification work [Smékal and Friedman, 2023, Koudahl et al., 2023].

fig. 7 maps the three scientific tracks — the analytical free-energy oracle, the on-policy pymdp student, and the formal/publication track (Lean, sheaf composition, provenance) — to 30 required pipeline tracks and 33 composable fragment types; the validation-gate registry itself indexes 27 gates.

Ontology-facing symbols are checked per model: the Bernoulli toy binds `pi1`, `pi2`, `J`, `gamma`, and `q_joint` — the teacher/student marginals, the coupling, the precision/temperature, and the joint posterior whose entanglement is the distillation signal — while the SI TMaze binds location/reward-location state factors, location/outcome/cue observations, q_π, first-action marginals, belief entropy, and SI tree evidence to **HiddenState**, **ObservationLikelihood**, **PolicyPosterior**, and **BeliefEntropy** terms (fig. 10, sec. 6).

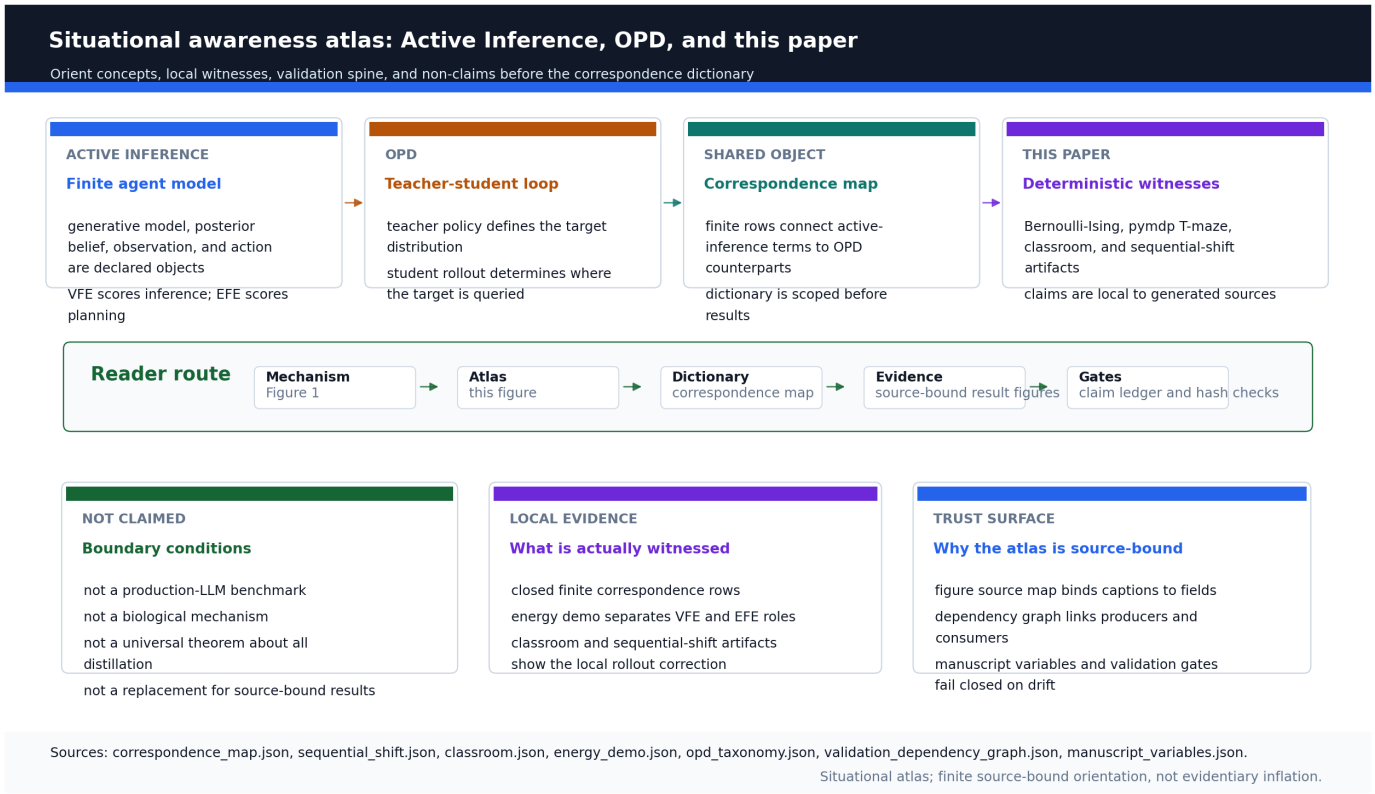


Figure 5: Source-bound situational-awareness atlas for the early manuscript. The figure is a finite orientation atlas, not a metric dashboard: it separates Active Inference primitives, OPD machinery, the correspondence dictionary, deterministic local witnesses, and explicit non-claims before the detailed correspondence map. The bottom panels make the scope guardrail visible: this is not a production-LLM benchmark, no biological mechanism is claimed, and it is not a universal theorem; it points to local deterministic artifacts and validation gates that later figures quantify or audit. Sources: `output/data/firstprinciples/correspondence_map.json`, `output/data/firstprinciples/sequential_shift.json`, `output/data/firstprinciples/classroom.json`, `output/data/firstprinciples/energy_demo.json`, `output/data/firstprinciples/opd_taxonomy.json`, `output/data/validation_dependency_graph.json`, and `output/data/manuscript_variables.json`.

4.1.1 Ontology bindings

- `expected_free_energy` → **ExpectedFreeEnergy**
- `location` → **HiddenState**
- `observation` → **ObservationLikelihood**
- `policy` → **PolicyPosterior**

Reader dictionary: active inference <-> shared object <-> on-policy distillation (26 rows)

Active inference		shared formal object		On-policy distillation
Kullback-Leibler information divergence	⟷	*directional KL between two categorical laws*	⟷	Teacher/student distribution discrepancy
Mean-field variational family	⟷	*tractable approximate posterior family*	⟷	Student policy family pi_S(y | x)
Generative model p(o, s)	⟷	*intractable target distribution*	⟷	Teacher policy pi_T(y | x, I)
Approximate posterior q(s)	⟷	*variational distribution being optimised*	⟷	Student policy pi_S(y | x)
Variational free energy F = D_KL(q || p) - log p(o)	⟷	*reverse Kullback-Leibler divergence*	⟷	Per-token reverse-KL distillation loss
Active sampling to minimise F	⟷	*the posterior generates its own observations*	⟷	On-policy student rollouts
Epistemic value (information gain)	⟷	*expected-free-energy exploration term*	⟷	Teacher signal on novel student states
Expected free energy risk/ambiguity split	⟷	*decomposition of future loss into preference risk and uncertainty/ambiguity terms*	⟷	Mode-seeking versus coverage-preserving distillation pressure
Expected free energy action selection (active sampling)	⟷	*EFE-minimising data-collection policy; epistemic value equals the closeable student-teacher gap*	⟷	On-policy choice of which states to roll out and distil on
Pragmatic value (prior preference)	⟷	*exp(R / beta) tilt of the prior*	⟷	Reward-tilted distillation target
Control-as-inference posterior	⟷	*Boltzmann policy posterior proportional to pi_ref exp(R / beta)*	⟷	KL-constrained RLHF / OPD target
RL-as-inference caveat	⟷	*modelled optimality variable plus approximation assumptions*	⟷	When reward-tilted posteriors stop matching deployable policy learning
Maximum-entropy trajectory posterior	⟷	*path distribution proportional to reference dynamics times exp(return / beta)*	⟷	Student rollout distribution under reward tilt
Direct preference posterior	⟷	*reference policy tilted by an implicit reward model*	⟷	DPO-style preference target for distillation
Markov blanket	⟷	*conditional-independence boundary*	⟷	Teacher/student context asymmetry
Predictive-coding hierarchy	⟷	*top-down target and bottom-up residual*	⟷	Teacher predictions and student prediction-error correction
Privileged sensory access	⟷	*conditioning variable absent at inference*	⟷	Privileged information I (hint, trace, feedback)
Context-conditioned teacher	⟷	*training-time conditioning context*	⟷	Teacher scored with instructions, scratchpads, demonstrations, or verified traces removed at deployment
Transferable versus shortcut privilege	⟷	*evidence mask over token positions supported by privileged context*	⟷	Evidence-localized OPD updates
Perception-action loop	⟷	*iterated variational optimisation*	⟷	generate rollout -> distill -> update policy
Sophisticated inference (beliefs about beliefs)	⟷	*recursive expected free energy*	⟷	Teacher conditioned on verified student traces
Precision / inverse temperature gamma	⟷	*confidence weighting of the target*	⟷	Distillation temperature beta
Divergence direction and estimator reliability	⟷	*choice of projection geometry and gradient estimator under distribution mismatch*	⟷	Forward-KL, reverse-KL, skew-KL, trust-region, and control-variate OPD variants
Long-horizon teacher reliability	⟷	*state-dependent trust in dense teacher targets*	⟷	Step-wise, long-context, and agentic OPD filters
Homeostasis (preserving priors while adapting)	⟷	*free-energy minimisation against a moving model*	⟷	Continual learning without forgetting (SDFT)
Sheaf local-to-global gluing	⟷	*finite consistency law over local evidence fragments*	⟷	Manuscript artifact contract and verifier composition

Source: output/data/firstprinciples/correspondence_map.json; every row is machine-validated.

Figure 6: The audited correspondence dictionary, rendered in full: all 26 machine-validated rows pairing an active inference construct (left) with its on-policy distillation counterpart (right) through the shared formal object both instantiate (center). This is the paper's thesis as a single picture: each row is a checked entry in `firstprinciples.mapping.CORRESPONDENCES` (no empty cells, unique keys), not a rhetorical analogy. The full table with per-row notes appears in the appendix. Source: `output/data/firstprinciples/correspondence_map.json`.

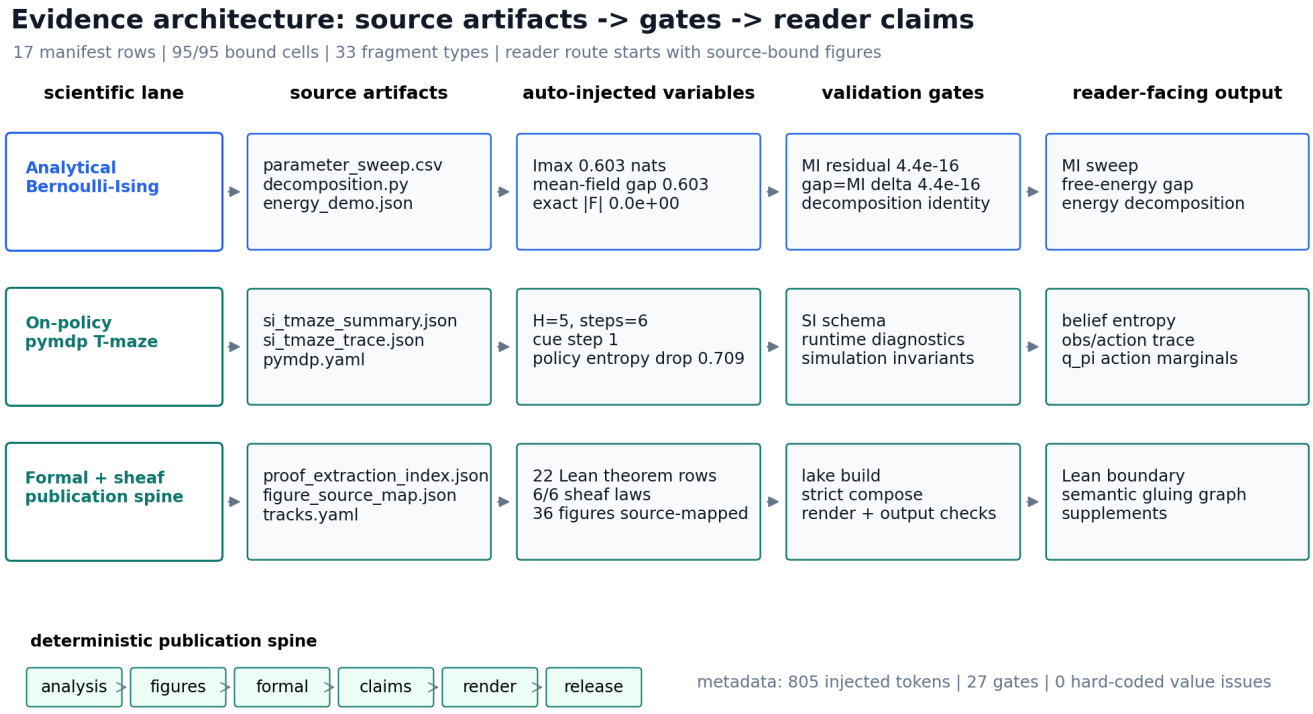
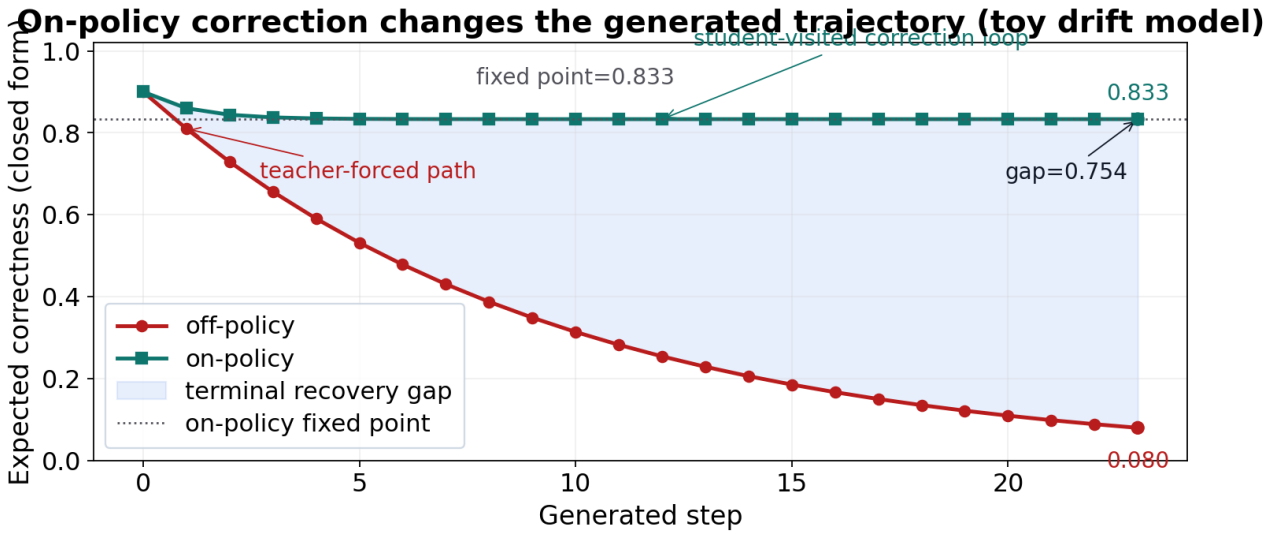


Figure 7: Evidence architecture for the manuscript. Each scientific lane exposes the complete chain from source artifacts to injected variables, validation gates, and reader-facing outputs: the analytical lane carries the mutual-information and free-energy gap values; the pymdp lane carries cue timing and policy-entropy diagnostics; the formal lane carries Lean theorem extraction, sheaf-law verification, figure provenance, and supplemental material. The bottom spine shows the deterministic publication sequence from analysis through release, with 805 injected manuscript tokens, 27 validation gates, and 0 hard-coded variable issues in the generated audit.



Source: output/data/firstprinciples/exposure_bias_demo.json

Figure 8: Toy model of exposure bias and its on-policy remedy: expected correctness — the closed-form per-step probability that the student emits the teacher-correct token under the two-state drift model — over generated steps for an off-policy student (trained on teacher-visited states only) versus an on-policy student (corrected on its own rollouts). The off-policy curve decays as compounding errors push the student into states the teacher never demonstrated – off-policy final correctness 0.080 – while the on-policy curve stabilizes at 0.833, leaving a terminal recovery gap of 0.754 (values from `output/data/firstprinciples/exposure_bias_demo.json`). Both curves are deterministic closed forms, so no uncertainty intervals apply. This is the motivating argument for on-policy distillation — with the caveat that the empirical severity of exposure bias is task-dependent — and mirrors active inference's insistence on evaluating beliefs along the agent's own visited trajectory rather than a fixed reference distribution.

Methods

5 Teacher and student coupling: the analytical model

We instantiate a minimal **K=2 Bernoulli / Ising** coupling as the analytical core of the teacher-student correspondence: it is the smallest model in which the privileged variable a teacher conditions on and the answer a student must emit are entangled by a single tunable coupling. The entangled joint eq. 1 places the teacher's privileged variable and the answer in one distribution governed by the coupling, exactly as a privileged teacher policy $\pi_T(y \mid x, I)$ binds the hint I to its output while the student family $\pi_S(y \mid x)$ sees only x. This is the finite probabilistic analogue of learning using privileged information [Vapnik and Vashist, 2009], of the distillation/privileged-information bridge [Lopez-Paz et al., 2016], of privileged ERM capacity caveats [Sharoni and Sabato, 2023], and of context/privileged OPD methods that must separate transferable privilege from shortcut features [Snell et al., 2022, Ye et al., 2026, Lazaridis et al., 2026, Liu et al., 2026a], with the Markov-blanket idea used only as a statistical screening boundary rather than as a predictor of numerical entropy gaps [Kirchhoff et al., 2018].

The closed-form mutual information $I(\lambda)$ is therefore the teacher-student mutual information - the bits the privileged channel injects that a factorised, mean-field student cannot recover - and the coupling λ is the dial that sets how privileged the teacher actually is. Mutual-information distillation work motivates why this channel view matters beyond the toy, but this manuscript uses MI only as a closed-form oracle [Shrivastava et al., 2023]. The free energy of fitting an approximate posterior to this joint uses the same directional information divergence introduced by Kullback and Leibler [Kullback and Leibler, 1951] and the same tractable-family move as variational inference in graphical models [Jordan et al., 1999, Blei et al., 2017]. In the active-inference sense [Friston et al., 2006, 2009, Friston, 2010, Parr and Friston, 2019, Parr et al., 2022], minimising $F = D_{\mathrm{KL}}(q \,\|\, p(s \mid o)) - \log p(o)$ over the student family — the KL target being the exact posterior $p(s \mid o) \propto p(o, s)$ of the declared generative model — is the per-token reverse-KL distillation loss in this finite variational family. Its concentration behaviour is support-, parameterisation-, and optimization-dependent rather than a universal LLM law [Hernández-Lobato et al., 2016, Ke et al., 2019, Wu et al., 2024, GX-Chen et al., 2025]. The mode-covering forward KL recovers the supervised-fine-tuning limit on teacher-generated data [Buciluǎ et al., 2006, Hinton et al., 2015, Kim and Rush, 2016]. Skew, entropy-aware, contrastive, and hybrid KL variants are therefore not separate objectives in this toy; they are alternate points in the same divergence-direction design space [Ko et al., 2024, 2025, Jin et al., 2026, Zhu et al., 2026b]. The same coupling thus interpolates the divergence families catalogued across the on-policy distillation landscape [Liu, 2026, Song and Zheng, 2026], and the entangled-versus-factorised gap in $I(\lambda)$ is the information a model leaves on the table when it distils toward itself conditioned on privileged context rather than the unconditioned family [Zhao et al., 2026, Liu et al., 2026e]. The free-energy terminology is scoped to these finite variational calculations in the sense of mathematical reviews of the free-energy principle [Buckley et al., 2017], and we read this Bernoulli-Ising oracle strictly as a minimal-model demonstration of the correspondence - not a claim about production language

models. The measured MI curve is presented once in the results sweep (fig. 15); this methods section supplies the equations, the exact $I(\lambda)$ recomputation contract, and the GNN round-trip identity check (fig. 10).

First-principles simulators. Beyond the closed-form oracle, five deterministic simulator families in `src/firstprinciples` stress each leg of the correspondence dynamically: a generalised-knowledge-distillation sweep that scores the same teacher signal under student- versus teacher-visitation measures, a variational expectation-maximisation loop whose free energy must descend under its exact clean E/M alternation, a diversity Pass-at-k temperature sweep that trades finite-sample commitment against coverage, an adaptive-divergence controller that interpolates between reverse- and forward-KL geometries, and a four-state/two-action sequential-shift family that compares teacher-forced train visitation with student-induced test visitation before and after deterministic on-policy correction. The sequential-shift family now includes a correction-dose sensitivity sweep, so the witness is not certified by a single chosen correction level alone. Each simulator is seedless and closed-form except where it consumes the pymdp classroom; their measured behaviour is reported alongside the T-maze results (sec. 10). Their role is to stress the variational correspondence, not to certify OPD optimization stability at scale; that stability question is left to the recent teacher-reliability, adaptive-exposure, freshness/asynchrony, self-generated-rationale, stepwise, and long-horizon OPD literature [Li et al., 2026, Luo et al., 2026, Han et al., 2026, Liu et al., 2026b, Chen et al., 2026, Zelikman et al., 2022, Zhong et al., 2026, Zhang et al., 2026, Tian et al., 2026].

Measured sweep grid points: 21.

In the distillation reading, the two binary streams π_1, π_2 are the teacher and student policies, and the coupling λ measures how strongly the teacher's privileged variable is tied to the answer the student must reproduce. The uncoupled baseline is the product measure $q_0(\pi_1, \pi_2) = q_1(\pi_1)q_2(\pi_2)$ with $q_1 = q_2 = (\frac{1}{2}, \frac{1}{2})$. This is the finite mean-field variational family: a tractable factorisation used to approximate an entangled target, with the approximation error measured by directional KL information [Kullback and Leibler, 1951, Jordan et al., 1999, Blei et al., 2017]. The entangled joint over the pair satisfies

$$q_\lambda(\pi_1, \pi_2) = Z_\lambda^{-1}\, q_0(\pi_1, \pi_2)\, \exp(\lambda J(\pi_1, \pi_2)), \tag{1}$$

with partition function Z_λ and symmetric Ising coupling $J(\pi_1, \pi_2) = 1$ on agreement and 0 otherwise. Reading q_λ as the generative model $p(o, s)$ that a tractable student must approximate, λ controls exactly the teacher–student dependence that on-policy distillation exists to transfer: at $\lambda = 0$ the student gains nothing from the teacher, while at large λ the teacher's privileged information I is fully informative about the target. This is the minimal model of the coupling that reverse-KL distillation objectives [Gu et al., 2024, Agarwal et al., 2024], skew/hybrid variants [Ko et al., 2024, Zhu et al., 2026b], and their self-distillation descendants [Zhao et al., 2026, Liu et al., 2026e] are built to exploit. Let $\sigma(\lambda) = q_\lambda(\pi_1 = \pi_2)$ be the probability that the two streams agree (the diagonal mass of the 2×2 joint) — equivalently, the probability that an on-policy student rollout matches the privileged teacher; by symmetry both marginals are uniform. With binary entropy $H_b(p) = -p \log p - (1-p) \log(1-p)$ in nats, the joint entropy is $H(q_\lambda) = \log 2 + H_b(\sigma(\lambda))$ while each marginal contributes $\log 2$, so the teacher–student mutual information is

$$I(\lambda) = \sum_k H(q_k) - H(q_\lambda) = \log 2 - H_b(\sigma(\lambda)), \tag{2}$$

eq. 2 vanishes at $\lambda = 0$ ($\sigma = \frac{1}{2}$, independent streams — the teacher conveys no privileged signal, the SFT-style off-policy limit [Hinton et al., 2015]) and saturates at $\log 2$ as $\lambda \to \infty$ ($\sigma \to 1$, perfectly entangled — the teacher fully determines the student target, the self-distillation limit). $I(\lambda)$ is therefore an interpretable ceiling for this toy binary coupling — the mutual information between the two streams, read as the epistemic value of teacher feedback on student-generated states — and a finite reference scale for the reverse-KL classroom signal reported by the companion executable demonstration; we do not claim a general communication-theoretic bound beyond this construction. The reward-tilted companion artifact uses the same normalised target form as control-estimation duality, trajectory inference, maximum-entropy IRL, control-as-inference, maximum-entropy RL, DPO/RLHF, and KL-constrained preference fine-tuning, $\pi^*(y \mid x) \propto \pi_{\text{ref}}(y \mid x) \exp(R(y)/\beta)$ [Todorov, 2008, Toussaint, 2009, Ziebart et al., 2008, Levine, 2018, O'Donoghue et al., 2020, Millidge et al., 2020a,b, Tschantz et al., 2020b, Haarnoja et al., 2018, Ziegler et al., 2019, Rafailov et al., 2023]. These claims are limited to this analytical model and its companion artifacts; they are a faithful minimal-model demonstration of the correspondence, not a measurement on production LLMs or a claim that every reinforcement-learning algorithm is literally probabilistic inference. These symbols are the rows of `analytical_assumption_index.json`, so the derivation is auditable rather than asserted.

The same closed forms supply a complementary observable: the conditional policy entropy

$$H(\pi_2 \mid \pi_1) = H(q_\lambda) - \log 2 = H_b(\sigma(\lambda)), \tag{3}$$

the residual uncertainty about one stream after observing the other. eq. 3 carries the exact complement identity $I(\lambda) + H(\pi_2 \mid \pi_1) = \log 2$, partitioning each binary decision into what teacher feedback resolves (the epistemic value transferred) and what it leaves open. In the distillation reading this is the per-decision uncertainty that remains *after* the student has absorbed the teacher signal — the irreducible part no objective in the divergence family can transfer at that coupling. Every observable in this section is checked along two genuinely independent routes: the literal analytic expressions above (σ, tanh, H_b) against a partition-function enumeration of the exact 2×2 joint, across all 105 sweep rows with maximum absolute residual 4.6e-16 nats under a 10^{-12} validator tolerance; a perturbed row fails the gate even if the stored summary is left untouched.

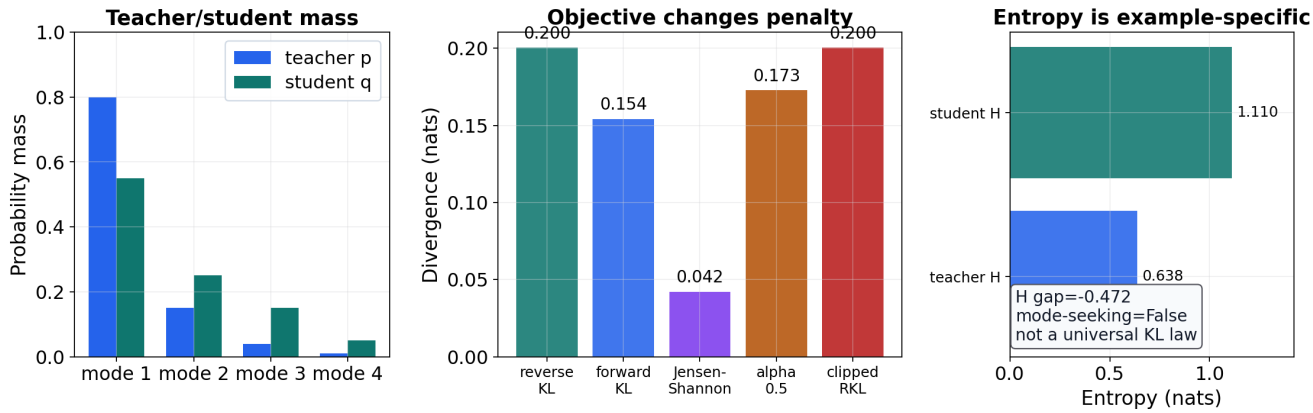
Proposition (scoped correspondence). *Assume* (A1) finite state, observation, and policy spaces as declared in the state-space catalog; (A2) the generative model is the explicitly constructed object — the entangled joint q_λ here, or the pymdp T-maze generative model in sec. 6; (A3) the student/posterior family is the declared tractable family (the mean-field product family here; the categorical pymdp posterior there); (A4) the student is evaluated on its own realized rollouts. *Then*: (i) **proved in closed form** — the per-decision reverse-KL distillation loss equals the variational free energy up to the evidence constant, $F = D_{\text{KL}}(q \,\|\, p(s \mid o)) - \log p(o)$ with KL target the exact posterior $p(s \mid o) \propto p(o, s)$, an algebraic identity of the declared objects; (ii) **proved in closed form, verified two-route**

— the information identities eq. 2 and eq. 3, with both derivation routes agreeing to 4.6e-16 nats; (iii) **demonstrated numerically** — gradient descent on the reverse-KL objective and on the variational free energy drive the same student to the same posterior (maximum absolute disagreement 3.6e-08, sec. 9); (iv) **interpretive** — reading on-policy rollouts as active sampling, and differential cue reliability as privileged information, is a correspondence built on (i)–(iii), not an additional theorem, and planning/action in the pymdp witness is selected by *expected* free energy rather than by the realized-rollout objective in (i). Outside (A1)–(A4) — sequence models, learned families, production-scale distillation — the correspondence is a structured analogy whose limits sec. 11 states explicitly.

The analytical track writes a parameter sweep comparing closed-form mutual information with an independent exact recomputation of it (via total correlation) across $\lambda \in [0, 4]$ on 21 grid points (sec. 8, fig. 15).

The `assumption_index` fragment makes the analytical equations inspectable as a generated artifact instead of relying on prose labels. `output/data/analytical_assumption_index.json` indexes 7 finite-model equation identifiers and 7 rows; the hydrated pass flag is `true`.

The index is deliberately narrow. It covers the Bernoulli-Ising toy equations, their finite binary state assumptions, and the generated artifacts that test the same symbols. Any missing equation identifier or empty assumption list fails the toy-sweep validation gate.



Source: output/data/firstprinciples/divergence_demo.json; realized behavior depends on support and optimization.

Figure 9: Divergence geometry for the teacher/student categorical toy. Left: teacher and student policy mass over four action modes for the fixed illustrative pair. Middle: the same teacher–student pair scored under five divergences – reverse KL 0.200 nats, forward KL 0.154 nats, Jensen-Shannon 0.042 nats, alpha-divergence 0.173 nats, and clipped reverse KL 0.200 nats (all from `output/data/firstprinciples/divergence_demo.json`). Right: the entropy panel shows the student entropy exceeds the teacher entropy, so this fixed illustrative pair is mode-covering (the artifact's mode-seeking flag is false), illustrating objective geometry rather than asserting a universal KL outcome. The spread across measures shows that the choice of distillation objective is not neutral: forward KL, reverse KL, alpha divergence, and clipping weight support mismatch differently, and the realized behavior remains support- and optimization-dependent. Implementation convention: inputs are projected onto the probability simplex, KL uses $0 \log 0 = 0$ and returns infinity on support violations rather than smoothing (no epsilon is added); the illustrative pair here has full support, so all values are finite and deterministic.

The Bernoulli toy is declared in `gnn/bernoulli_toy.gnn.md` (GNN v1.1), following the GNN notation role described by Smekal and Friedman [Smékal and Friedman, 2023]. fig. 10 links GNN variables to Active Inference Ontology terms bound in the analytical ontology fragment; round-trip parity is checked before render.

Measured MI and sweep artifacts in sec. 8 ground the same symbol map used in the concordance diagram.

5.0.1 Ontology bindings

- `E1` → **Stream1HabitPrior**
- `E2` → **Stream2HabitPrior**
- `J` → **CrossStreamCouplingPotential**
- `gamma` → **SophisticationWeight**
- `lam` → **EntanglementDeformationParameter**
- `pi1` → **Stream1PolicyVector**
- `pi2` → **Stream2PolicyVector**
- `q_joint` → **EntangledJointPosterior**

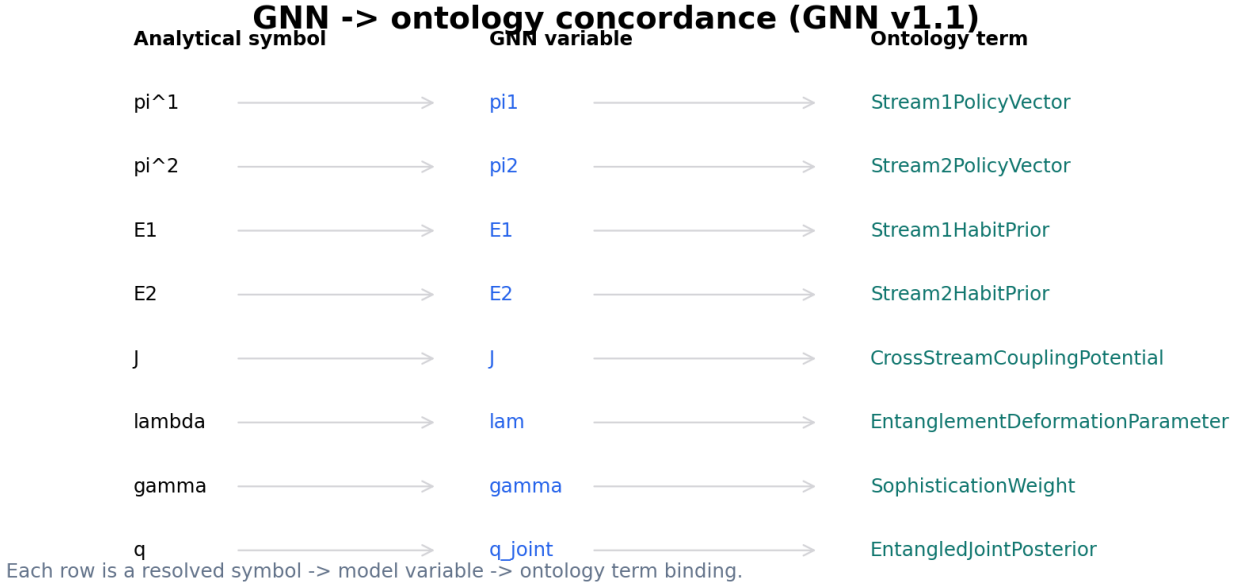


Figure 10: Concordance diagram aligning the analytical symbols of the Bernoulli–Ising toy with the generative-model variables declared in `bernoulli_toy.gnn.md` (GNN v1.1) and their corresponding Active Inference Ontology terms. The three-layer mapping covers all 8 expected variable–term pairs and checks that the equations in the manuscript, the executable GNN specification, and the shared ontology vocabulary use the same declared names. This naming concordance keeps the on-policy-distillation toy aligned with a GNN/ontology-scoped active-inference specification rather than relying on a loose analogy.

6 On-policy student: pymdp sophisticated inference

The on-policy student as sophisticated inference (planning horizon). Under our thesis, on-policy distillation is the student policy acting to minimise expected free energy: it generates its own observations through rollouts and is scored against a privileged target. This section documents the canonical **pymdp full TMaze sophisticated-inference validation profile** (fig. 11) as the executable minimal model of that on-policy student, with SI search horizon 5, rollout length 6, and Agent `policy_len = 1`. Sophisticated inference — beliefs about beliefs over the planning horizon — is precisely the structure invoked by self-distillation methods that condition a teacher on the student's own verified traces [Zhao et al., 2026, Liu et al., 2026e]; an agent that rolls out and then minimises expected free energy is the active-sampling counterpart of the student rollouts in on-policy distillation [Agarwal et al., 2024]. The discrete-state active-inference framing follows finite-POMDP treatments, comparisons with RL, scaling discussions, and tutorial literature [Da Costa et al., 2020, Sajid et al., 2021a, Smith et al., 2022, Parr et al., 2022, Tschantz et al., 2020a,b], with the expected-free-energy decomposition into epistemic and pragmatic terms supplying the information-gain-versus-reward split that distillation losses recapitulate [Millidge et al., 2021b, Friston et al., 2021a, Champion et al., 2024, Sajid et al., 2021b, de Vries et al., 2025]. We adopt this epistemic/pragmatic split as our reading; the precise decomposition of expected free energy is not canonical and varies by implementation across these treatments, so the variational free energy that scores the student's realized rollouts and the expected free energy that governs its counterfactual policy choice remain distinct objects here, and the information-gain-versus-reward partition we report is the one our pymdp witness instantiates rather than a framework-universal form. The implementation anchor is pymdp's `TMaze`, `Agent`, `si_policy_search`, and `rollout` APIs [Heins et al., 2022]. The configured profile is `full_tmaze_sophisticated_inference`; the canonical planner is `sophisticated_inference`.

The same situation in two frameworks. To make the correspondence concrete rather than analogical, we solve one scenario — two-state reward-location inference under an informative cue — in both the active-inference and the standard machine-learning idiom. On the active-inference side, the teacher is the exact Bayesian posterior $p(s \mid o)$, the unique minimiser of variational free energy. On the machine-learning side, a student categorical policy with logits θ (a softmax policy) is trained by gradient descent on the reverse-KL on-policy-distillation loss $D_{\mathrm{KL}}(\pi_S \| \pi_T)$ using `jax` automatic differentiation. The two procedures converge to the *same* distribution: the ML-distilled student reproduces the active-inference posterior to within 3.6e-08 (the artifact's frameworks-agree flag is true), and its variational free energy reaches 0.693 nats, matching the evidence bound $-\ln p(o) = 0.693$ nats. Minimising the reverse-KL distillation loss and minimising variational free energy therefore reach the *same* optimum here: the distillation run converges onto the active-inference posterior, executed by gradient descent rather than asserted by analogy [Levine, 2018, Penaloza et al., 2026a].

The generative process has 5 location states/actions, 2 reward-location states, and 3 observation modalities (location, outcome, cue). The cue modality carries the privileged information of the distillation correspondence: it is the hint, verified trace, or feedback channel available in training but not guaranteed at inference, and `cue_validity` is the strength of that privilege [Friston, 2013, Kirchhoff et al., 2018, Vapnik and Vashist, 2009, Lopez-Paz et al., 2016, Cai et al., 2024]. That makes the asymmetry operational: the teacher has privileged sensory access, whereas the on-policy student must act to sample the channel and is then evaluated on the trajectory it actually induced. The model/value audit in `output/data/si_tmaze_model_matrices.json` records A=[[5, 5], [3, 5, 2], [3, 5, 2]]; B=[[5, 5, 5], [2, 2, 1]], dependencies, preferences, deterministic D priors, reward condition 0, and cue validity 0.95 (fig. 12). Per-step trace

records include q_π, marginal first-action probabilities, selected action names, modality-specific observations, belief entropy, and SI tree metadata — the on-policy trajectory the student writes and is then scored against.

Graph-world artifacts are deterministic extension outputs declared in `tracks.yaml extension_tracks.graph_world`. For the reference workflow, see sec. 3; measured rollouts appear in sec. 10.

Mean belief entropy across recorded timesteps: 0.1841 nats. Because the cue is informative, the student that acts to observe it drives its posterior entropy down — the epistemic value of seeking privileged information made quantitative. The initial q_π first-action marginal assigns probability 0.545 to the cue-directed action, which is the canonical first-action argmax under the configured SI search: the on-policy student elects, from its own beliefs, to sample the privileged channel first.

The comparison artifact `output/data/si_policy_comparison.json` runs the canonical SI planner alongside a vanilla pymdp planner as validation rows only; the vanilla planner is marked `comparison_only` and never replaces the canonical summary. It records 2 deterministic comparison rows, complete-grid flag 1, and 1 goal-reaching rows under the same full TMaze transition model. We read this contrast as the difference between an agent that minimises expected free energy on-policy and a myopic baseline. That contrast is the active-inference image of the induced-distribution issue from behavioral cloning, interactive imitation learning, sequence generation, policy distillation, and LLM distillation: the learner must be trained on states it actually causes, not only on teacher-generated states [Pomerleau, 1989, Ross and Bagnell, 2010, Ross et al., 2011, Sun et al., 2017, Bengio et al., 2015, Arora et al., 2022, Rohatgi et al., 2025, Pozzi et al., 2025, Hinton et al., 2015, Kim and Rush, 2016, Rusu et al., 2016, Czarnecki et al., 2019, Agarwal et al., 2024].

Agent construction and rollout diagnostics are audited in `output/reports/pymdp_runtime_diagnostics.json`: 2 constructions, 2 known third-party JAX static-array warnings, 39 known SI tree max-node diagnostics, and 0 unexpected warnings. Policy posterior evidence is written separately to `output/data/pymdp_policy_posterior_grid.json` with 14 rows and normalized-posterior flag 1.

The graph-world extension is deterministic: `simulate_si_graph_world.py` writes `si_graph_world_summary.json` and `si_graph_world_trace.json` for a four-node graph-world path. The regenerated summary reports 4 nodes, 4 steps, and goal-reached flag 1. The topology-trace extension records 4 topology traces with agreement flag 1. As with the analytical toy, these are a minimal-model demonstration of the on-policy-distillation/active-inference correspondence — claims are limited to these pymdp models and artifacts, not to production LLM systems.

Given generative matrices A, B, C, D, pymdp computes state beliefs $q(s)$ and policy posterior rows q_π inside `rollout`. The canonical SI Agent uses `policy_len = 1`, while `si_policy_search` supplies the effective search horizon $H = 5$ (logged with `num_policies = 5` and tree metadata in the SI summary artifact; see sec. 10).

The default harness records belief entropy per step and q_π first-action probabilities from the sophisticated-inference rollout. Vanilla planning is retained only in `output/data/si_policy_comparison.json` as comparison evidence, not as a manuscript co-primary track.

SI artifacts (summary, trace, optional JSONL log) record the canonical full_tmaze_sophisticated_inference rollout: 6 rollout transitions, 7 recorded timesteps, 3 observation modalities, 7 q_π rows, 7 marginal first-action probability rows, and SI tree availability flag 1.

The `interop` fragment treats the GNN files, JSON views, and ontology bindings as a round-trip contract rather than parallel documentation. That places this manuscript's GNN use in the broader effort to make active-inference models diagrammatically specified and machine-checkable [Smékal and Friedman, 2023, Koudahl et al., 2023]. `output/data/interop_roundtrip_report.json` records 6 deterministic checks and reports lossless round-trip status `true`; the manuscript only claims losslessness from that generated flag.

The stricter lint artifacts are adjacent evidence, not new model claims: `output/data/gnn_roundtrip_report.json`, `output/reports/gnn_lint_report.json`, `output/data/ontology_alias_index.json`, and `output/data/ontology_profile_matrix.json` must agree before the interop row passes. A missing GNN variable, duplicate ontology alias, dropped JSON field, shape diff, or dtype diff is therefore a validation failure before rendering.

See `gnn/si_tmaze.gnn.md` for a GNN view of the full pymdp TMaze hidden-state factors, three observation modalities, q_π first-action posterior, belief entropy, and SI tree evidence with ontology bindings.

6.0.1 Ontology bindings

- `belief_entropy` → **BeliefEntropy**
- `first_action_prob` → **PolicyPosterior**
- `loc` → **HiddenState**
- `obs_cue` → **ObservationLikelihood**
- `obs_location` → **ObservationLikelihood**
- `obs_outcome` → **ObservationLikelihood**
- `q_pi` → **PolicyPosterior**
- `reward_loc` → **HiddenState**
- `si_tree_nodes` → **PolicyPosterior**

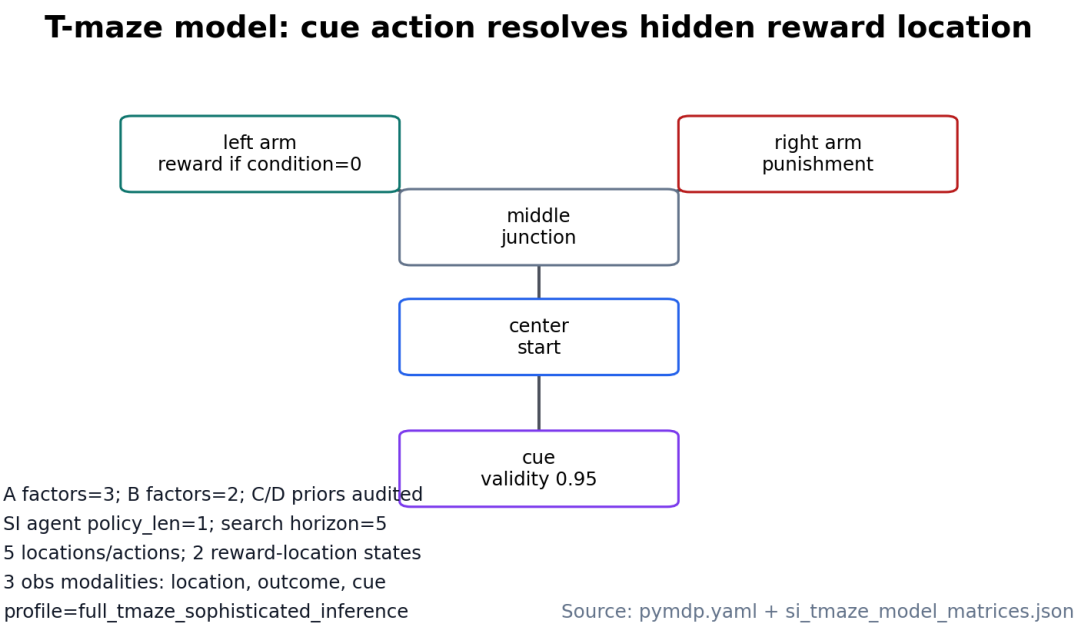


Figure 11: Schematic of pymdp's full TMaze generative process: 5 location states/actions, 2 reward-location states, three observation modalities, cue validity 0.95, and SI search horizon 5. The diagram lays out the hidden states, observations, and controllable transitions that define the teacher's task, including the cue arm whose validity determines how informative the epistemic action is. It fixes the world model in which the active-inference teacher is computed and, in turn, the task structure an on-policy student must operate within to be distilled faithfully.

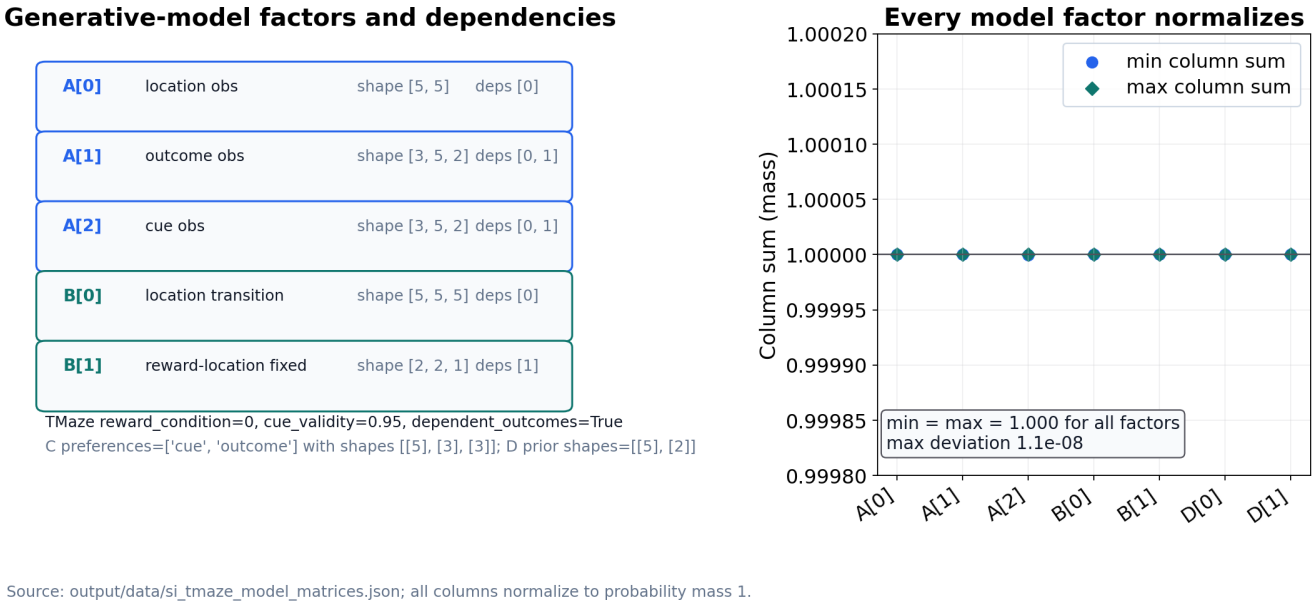
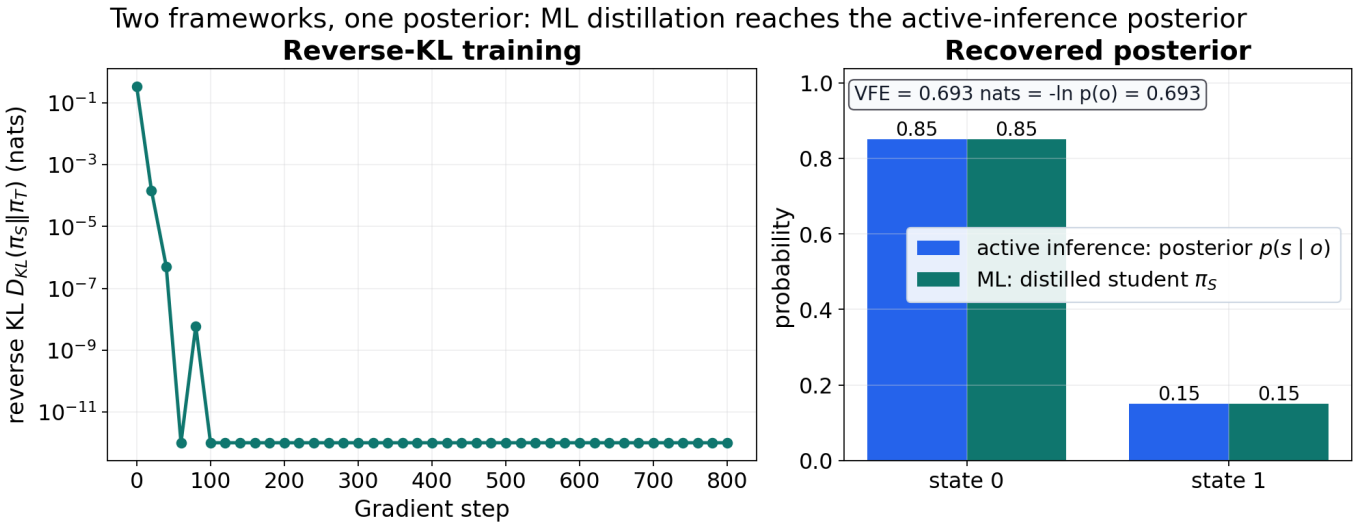


Figure 12: Full TMaze generative-model matrix and value audit. Left: the labeled A (likelihood) and B (transition) factors with shapes A=[[5, 5], [3, 5, 2], [3, 5, 2]]; B=[[5, 5, 5], [2, 2, 1]] and their state dependencies, alongside C preferences and D priors. Right: A, B, and D normalization checks confirming each conditional distribution sums to unit probability mass (from `output/data/si_tmaze_model_matrices.json`). The audit exposes the exact parameters that generate the teacher's behavior and verifies they are valid probabilities, so the active-inference policy being distilled rests on a well-formed generative model rather than an unchecked numerical artifact.



Source: output/data/firstprinciples/parallel_demo.json

Figure 13: One scenario solved in two frameworks. A standard machine-learning loop (jax automatic differentiation on the reverse-KL distillation loss) drives a student policy to the active-inference exact posterior $p(s \mid o)$: the per-step reverse KL decays toward zero (left panel, logarithmic y-axis) and the distilled student matches the posterior to within 3.6e-08 (right), its variational free energy reaching 0.693 nats – the evidence bound $-\ln p(o) = 0.693$. Optimization: full-batch gradient descent at learning rate 0.5, at most 800 steps with early stop at loss tolerance 1e-05; the run is fully deterministic (no sampling), so no uncertainty intervals apply. Minimising the reverse-KL distillation loss converges to the same minimiser as variational free energy, so the correspondence is executed rather than asserted. Source: `output/data/firstprinciples/parallel_demo.json`.

7 Machine-checked correspondence (Lean)

The Lean track is a boundary layer, not a proof of the full OPD=AI thesis. Its role is to make the finite assumptions used by the executable artifacts explicit: the centered Ising coupling witness, T-maze reachability and absorbing-goal witnesses, graph-world reachability witnesses, finite policy enumeration, finite belief-weight normalization, finite policy-posterior normalization, the positive-horizon witness for sophisticated inference, and the finite-channel chain-rule skeleton for the mutual-information complement identity. These are compiled by `lake build`, and their proofs are audited at the kernel level by an axiom gate that elaborates the source directly – so a `sorry` or a wrong definition cannot hide behind a cached build – and then extracted into `output/data/proof_extraction_index.json`, where 22 theorem rows must match the Lean theorem inventory and constructive-token status is `true`. The validation gates also fail if a theorem statement is dropped, if `sorry`, `axiom`, or `native_decide` appears, or if the generated theorem index diverges from `output/reports/lean_theorem_inventory.json`.

The central correspondence still lives in the analytical and simulation artifacts: the reverse-KL/variational-free-energy analogy is derived in sec. 5, while the positive-horizon on-policy sampling interpretation is exercised in the pymdp T-maze and graph-world traces. The Lean layer certifies the small finite boundaries those arguments depend on, including the parameterized `tmaze_goal_absorbing` witness, the `sophisticated_requires_horizon` witness, and the `mi_chain_rule` skeleton — bracketed by a `cue_closes_gap` positive witness and a `pragmatic_leaves_gap` negative control — that machine-checks the *algebraic structure* of the mutual-information complement identity the active-selection result rests on — `I(o;r) + E_o[H(r\mid o)] = H(r)` over the integers. The two flanking theorems pin the active-selection endpoints exactly: a cue that drives the residual to zero provably transfers the *full* prior entropy ($I = H(r)$, the cue-visiting policy), while any strictly positive residual provably transfers strictly *less* ($I < H(r)$, the pragmatic-only policy that leaves the distillation gap open), so the skeleton is non-vacuous in both directions. The same module proves the skeleton's structural shape: the conditional-entropy fold is additive over channel concatenation (`expectedCondEntropy_append`), the mutual information is bounded $0 \le I(o;r) \le H(r)$ under non-negative residuals (`mi_bounded` — the integer image of "epistemic value is bounded by the prior entropy", the property the active-selection sweep exhibits as epistemic value ranges over 0 to $H(r)$), it is antitone in the residual (`mi_antitone`), and a residual equal to the prior entropy transfers nothing (`blind_channel`, the dual of `cue_closes_gap`). This is deliberately the finite chain-rule skeleton, **not** the real-valued entropy identity `I + H_b(\sigma) = \log 2`, which remains the two-route numerical witness in sec. 5 (the Lean toolchain here ships without Mathlib's real-valued entropy). It does *not* prove a general theorem about pymdp's q_π posterior, production language models, or all sophisticated-inference planners. The paper therefore treats Lean as a checked finite-witness interface: it binds the toy state machines and horizon assumptions to named theorem rows, then the Python gates bind those theorem rows to generated artifacts and manuscript claims.

That interface is intentionally redundant with the non-Lean finite checks rather than a replacement for them. `output/reports/model_checking_witnesses.json` contributes 12 exhaustive toy witnesses with all-pass status `true`; `output/data/theorem_traceability_matrix.json` contributes 22 linked theorem rows; and the Lean graph-world inventory witnesses 4 generated topology ids with all-topologies-witnessed flag `true`. fig. 14 summarizes this proved-versus-deferred boundary without duplicating proof scripts in prose.

The Lean `SophisticatedInference` boundary module declares the finite planning-horizon witness used to mirror the pymdp SI

Module	Kind	Name	Status
OPD.BernoulliToy	def	isingCouplingEntries	proved
OPD.BernoulliToy	def	isingCouplingSum	proved
OPD.BernoulliToy	theorem	ising_coupling_sum_zero	proved
OPD.InformationIdentity	def	expectedCondEntropy	proved
OPD.InformationIdentity	def	mutualInformation	proved
OPD.InformationIdentity	theorem	mi_chain_rule	proved
OPD.InformationIdentity	def	cueResolvesObs	proved
OPD.InformationIdentity	theorem	cue_closes_gap	proved
OPD.InformationIdentity	theorem	pragmatic_leaves_gap	proved
OPD.InformationIdentity	theorem	foldl_add_acc	proved
OPD.InformationIdentity	theorem	expectedCondEntropy_append	proved
OPD.InformationIdentity	theorem	expectedCondEntropy_nonneg	proved
OPD.InformationIdentity	theorem	mi_bounded	proved
OPD.InformationIdentity	theorem	mi_antitone	proved
OPD.InformationIdentity	theorem	mi_antitone_strict	proved
OPD.InformationIdentity	theorem	blind_channel	proved
OPD.InformationIdentity	def	blindObs	proved
OPD.InformationIdentity	theorem	blind_witness	proved
OPD.SophisticatedInference	def	defaultPolicyLen	proved
OPD.SophisticatedInference	theorem	sophisticated_requires_horizon	proved
OPD.SophisticatedInference	def	tmazeStep	proved
OPD.SophisticatedInference	theorem	tmaze_two_forward_steps_reach goal	proved
OPD.SophisticatedInference	theorem	tmaze_goal_absorbing	proved
OPD.SophisticatedInference	def	graphWorldStep	proved
OPD.SophisticatedInference	theorem	graph_world_three_steps_reach goal	proved
OPD.SophisticatedInference	def	branchGraphWorldStep	proved
OPD.SophisticatedInference	theorem	branch_graph_world_three_steps reach_goal	proved
OPD.SophisticatedInference	def	loopGraphWorldStep	proved
OPD.SophisticatedInference	theorem	loop_graph_world_four_steps reach_goal	proved
OPD.SophisticatedInference	def	diamondGraphWorldStep	proved
OPD.SophisticatedInference	theorem	diamond_graph_world_four_steps reach_goal	proved
OPD.SophisticatedInference	def	finitePolicies	proved
OPD.SophisticatedInference	theorem	policy_enumeration_contains forward	proved
OPD.SophisticatedInference	def	twoStateBeliefWeights	proved
OPD.SophisticatedInference	theorem	two_state_belief_weights_sum to_two	proved
OPD.SophisticatedInference	def	twoPolicyPosteriorWeights	proved
OPD.SophisticatedInference	theorem	two_policy_posterior_weights sum_to_two	proved

Figure 14: Lean formalization boundary: a table of modules, declaration kinds, names, and proved-versus-sorry status under `lean/OnPolicyDistillation/`, each row a witness checked by `lake build`. Proved rows mark the finite boundary claims in this inventory that are machine-verified, while any sorry row honestly demarcates the edge of what is formally established. The figure makes the trust boundary explicit: the compiled core is the declared finite witness set, not a general proof about all OPD or active-inference systems.

search horizon 5, alongside finite T-maze boundary witnesses such as `tmaze_two_forward_steps_reach_goal` and `tmaze_goal_absorbing`. It also contains constructive finite witnesses for graph-world reachability, finite policy enumeration, belief weights, and policy-posterior weights. These theorems formalize small finite boundaries shared with generated artifacts; they do *not* prove that the pymdp q_π posterior is a general model of sophisticated inference. The companion `InformationIdentity` module adds the finite-channel chain-rule skeleton over `Int`: `mi_chain_rule` (the complement identity for any prior and any finite observation list), the `cue_closes_gap` / `blind_channel` endpoints (full versus zero transfer), the `pragmatic_leaves_gap` negative control, and the structural properties `expectedCondEntropy_append` (additivity), `mi_bounded` ($0 \le I \le H(r)$), and `mi_antitone` (antitone in the residual) — all proved by `omega`, `simp`, and induction with no `Real.log` and no Mathlib dependency. Axioms are audited with `#print axioms` (the gate whitelists only `propext`, `Classical.choice`, `Quot.sound`) over *every* theorem discovered from source, so a new theorem cannot escape the audit; see the Lean track gate.

Build via `lake build` under `lean/`.

The `model_checking` fragment complements Lean with finite exhaustive witnesses. `output/reports/model_checking_witnesses.json` records 12 toy-state witnesses and reports `true` only when no counterexample is found in the enumerated state/action space.

This is deliberately narrower than a semantic proof of all Active Inference programs. It checks the finite T-maze and graph-world boundary objects used by this manuscript and exposes the witness inventory to the same artifact and claim gates as the Lean theorem inventory. The Lean graph-world inventory witnesses 4 generated toy topology ids, with all-topologies-witnessed flag `true`; theorem traceability contributes 22 linked rows.

The `theorem_traceability` fragment binds Lean theorem inventory rows to finite model-checking witnesses, manuscript claims, and evidence fields. `output/data/theorem_traceability_matrix.json` records 22 traceability rows and passes only when every theorem row is linked (`true`).

7.0.1 Proof extraction track

The `proof_extraction` track extracts Lean theorem statements and proof-source metadata into `output/data/proof_extraction_index.json`. The index currently contains 22 extracted theorem rows, with constructive-token status `true`.

The extracted rows are checked against `output/reports/lean_theorem_inventory.json` before the manuscript can render. This catches a false-green case where `lake build` passes but a theorem silently falls out of the generated proof index; the gate requires the theorem inventory and extracted proof rows to agree exactly.

Results

8 Teacher and student mutual information

Under the correspondence of this paper, the coupling strength λ is the degree to which the teacher's privileged variable is bound to the answer the student must produce: at $\lambda = 0$ the teacher's hint and the answer are independent, so there is nothing privileged to transfer, while as λ grows the teacher acquires a strictly informative channel that the on-policy student lacks at inference time. The mutual information $I(\lambda)$ is therefore exactly the teacher–student mutual information of the entangled joint, the upper bound on how much the privileged generative model $p(o, s)$ can communicate to the tractable posterior $q(s)$ being fit [Friston, 2010, Parr et al., 2022]. We sweep coupling strength λ on a grid of 21 points up to $\lambda_{\max} = 4$, tracing this coupling curve from the decoupled limit to maximal entanglement. Closed-form mutual information from eq. 1 is cross-checked against an independent exact recomputation via total correlation from the analytical module (sec. 5); both are deterministic (no sampling) and agree to within 4.4e-16 nats — machine precision, not exact zero.

The curve rises monotonically in λ: inside this finite joint, more coupling means more transferable information. That is the whole role of this result. It supplies a closed-form axis for teacher-student informativeness; it does not reproduce or explain the literature-reported Qwen/Thinking Machines production rows, which remain external context in the discussion [Qwen Team, 2025, Lu and Thinking Machines Lab, 2025]. The free-energy calculation in sec. 9 separates two cases that are easy to conflate. When the variational distribution is the exact coupled target, the target-relative free energy is zero up to 0 numerical residual. When the student is forced into the independent mean-field family at the same coupling, the penalty is the information gap 0.603 nats, matching mutual information to within 4.4e-16 nats. In this finite reverse-KL target-family calculation, the gap is what an on-policy student must close by using the teacher's privileged samples rather than fitting an unconditional factorised family [Gu et al., 2024, Agarwal et al., 2024, Zhao et al., 2026, Liu et al., 2026e]. The forward direction, by contrast, recovers the teacher-data supervised-fine-tuning limit with its attendant exposure bias [Hinton et al., 2015]. The alpha/f-divergence and KL-geometry literature is the guardrail: the toy demonstrates a support-aware variational role, not a universal statement that one divergence direction always yields one LLM behavior [Hernández-Lobato et al., 2016, Ke et al., 2019, Wu et al., 2024]. The classroom simulation makes the same local motif executable: a privileged teacher at high cue validity transfers a posterior-sharpness advantage to an on-policy student through a reverse-KL signal, and the analytical $I(\lambda)$ curve is a closed-form ceiling for this specific toy channel. These results are a minimal-model demonstration of the correspondence between the variational and distillation objectives, not a claim about production language models; they hold for the analytical toy and its artifacts.

The sweep reuses the entangled joint defined in eq. 1 (sec. 5). Mutual information $I(\lambda) = \log 2 - H_b(\sigma(\lambda))$ is evaluated on the same λ grid as the analytical oracle and its independent exact recomputation.

Both estimators are deterministic (no sampling, no RNG) and are evaluated on the same λ grid as the closed-form sweep (sec. 5, fig. 15).

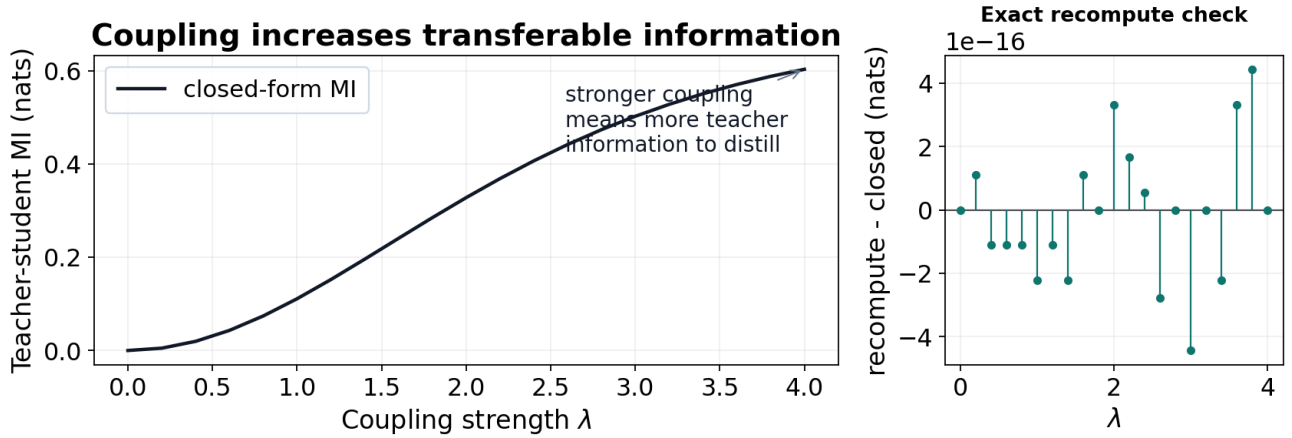


Figure 15: Mutual information between the two coupled spins as a function of coupling strength. Left: closed-form $I(\lambda)$ for the symmetric Bernoulli-Ising toy across 21 grid points up to $\lambda_{\max} = 4$, rising monotonically toward a grid maximum of 0.6031 nats as the spins become maximally correlated. This is the minimal model of the teacher–student coupling that on-policy distillation must transmit: the analytic information content the student policy is asked to absorb from the teacher. Right: the independent exact recomputation via total correlation is shown as recompute-minus-closed-form residuals rather than as a second overplotted curve; because both estimators are deterministic (no sampling), the maximum residual of 4.4e-16 nats (machine precision) is a cross-implementation agreement check confirming the analytic information measure is reproducible to machine precision.

9 Free-energy decomposition

Read as a distillation-objective landscape, the variational free energy of the student against the reward-tilted teacher target is evaluated along the same λ grid used for the MI sweep (fig. 16), where λ couples the teacher's privileged variable to the answer and thus sets how much teacher–student mutual information the objective must absorb. fig. 16 separates two objects that are easy to conflate. Against the *entangled* target p_λ — the analogue of the teacher policy $\pi_T(y \mid x, I)$ that carries privileged information I — the entangled posterior q_λ is evaluated against its own normalized target, so $F(q_\lambda; p_\lambda) = 0$ to numerical tolerance (maximum absolute value 0.0e+00 nats). That zero is not an absence of structure; it is the finite self-distillation limit [Zhao et al., 2026] where the tractable student family has exactly recovered the target it is fit against and the reverse-KL loss vanishes.

The decomposition implemented in `src/analytical/decomposition.py` then splits that zero into per-stream marginal free energies, a coupling-cost term, a coupling-prior term, and a total-correlation gain — the same epistemic/pragmatic ledger that expected-free-energy methods balance in active inference [Friston, 2010, Parr et al., 2022, Millidge et al., 2021b, Champion et al., 2024, Sajid et al., 2021b]. For the symmetric toy with uniform marginals, the coupling-prior term equals $-I(\lambda)$ and exactly cancels the total-correlation gain $+I(\lambda)$; the merged invariant suite checks this cancellation directly (16/16 pass). This is the precise sense in which teacher–student mutual information is the slack between the coupled target and the factorized family for this model.

The nonzero curve in fig. 16 is the free-energy gap against the *mean-field* independent prior q_0, which plays the role of the mode-covering, forward-KL SFT baseline [Hinton et al., 2015] that ignores the privileged coupling. Its minimum at $\lambda = 0$ occurs where the entangled posterior coincides with the factorized mean-field product; any $\lambda > 0$ raises the gap as coupling pulls the posterior away from that independent prior. At the configured sweep maximum the gap reaches 0.603 nats and equals mutual information up to 4.4e-16 nats, so the rising branch is the information gap the on-policy reverse-KL objective must close [Gu et al., 2024, Agarwal et al., 2024]. As the privileged variable becomes more diagnostic, a student that cannot condition on I pays an increasing cost, exactly the regime where on-policy self-distillation toward a teacher conditioned on verified traces is designed to operate [Liu et al., 2026e, Hübotter et al., 2026].

Saturation MI (grid maximum on the measured λ sweep): 0.6031 nats. This is the largest coupling attained on the measured λ grid, not the model's ceiling: as $\lambda \to \infty$ the mutual information saturates toward $\log 2 \approx 0.693$ nats (eq. 2), which is the information the privileged target can carry beyond the student's reach in this binary toy. As throughout, the claim is limited to this analytical toy as a faithful minimal demonstration of the correspondence, not an assertion about production LLMs.

9.0.1 Energy decompositions: VFE and EFE

To make the correspondence between the reverse-KL distillation loss and variational free energy (VFE) explicit, we evaluate the two energy ledgers of active inference on the same minimal model and report them in fig. 17. VFE admits the standard complexity-minus-accuracy reading [Friston, 2010, Parr et al., 2022]: $F = \underbrace{D_{\mathrm{KL}}[q(s) \,\|\, p(s)]}_{\text{complexity}} - \underbrace{\mathbb{E}_q[\ln p(o \mid s)]}_{\text{accuracy term}}$, equivalently the negative log-evidence offset by the posterior-to-true-posterior gap. In the prior-evaluated panel, the complexity term is 0.000 nats because $q(s) = p(s)$, so the KL cost of moving off the prior is zero. The accuracy term is -1.030 nats because it is an expected log likelihood and is negative under this observation model; therefore the plotted VFE is $0 - (-1.030)$, the positive surprisal term behind the reported log-evidence bound -0.693 nats. The point is not that energy disappears, but that the cost comes entirely from insufficient accuracy when the student remains at the prior. In distillation language, that is the unlearned teacher signal: minimizing F tightens the reverse KL between the tractable student $\pi_S(y \mid x)$ and the privileged target $\pi_T(y \mid x, I)$ up to the model-evidence constant $-\ln p(o)$ that is invariant to the student parameters [Levine, 2018, Da Costa et al., 2020]. This correspondence is an algebraic objective identity inside the declared finite categorical family — the variational reading of the divergence in the sense of the free-energy-principle mathematical review [Buckley et al., 2017] — so a shared objective form fixes the optimum here but does not carry over to the optimization dynamics or scaling of large-model distillation, where the same reverse-KL objective behaves in support-, parameterisation-, and optimization-dependent ways rather than as a universal law [Wu et al., 2024].

Expected free energy (EFE) extends this ledger forward over the student's own rollouts, which is where the *on-policy* character of distillation enters: the student generates the very observations it is then scored against, so the relevant energy is an expectation over self-generated trajectories rather than a fixed teacher-forced corpus [de Vries et al., 2025, Liu et al., 2026e, Hübotter et al., 2026]. EFE splits two equivalent ways. The risk-plus-ambiguity reading assigns 0.511 nats to risk (the divergence of predicted from preferred outcomes — the reward-tilting term that biases rollouts toward the verified target) and 0.423 nats to ambiguity (the expected observation entropy under the generative mapping). The epistemic-plus-pragmatic reading assigns 0.270 nats to the epistemic (information-gain) drive and -1.204 nats to the pragmatic (goal-seeking) drive. The pragmatic term is the active-inference image of reward-tilting in on-policy distillation, while the epistemic term is the active-sampling pressure that makes on-policy rollouts close the exposure-bias gap the off-policy SFT baseline leaves open [Qwen Team, 2025, Lu and Thinking Machines Lab, 2025, Zhao et al., 2026].

The full variational- and expected-free-energy decomposition is tabulated in the supplement (sec. 13).

The table makes the thesis quantitative on one model: VFE is the static distillation objective (complexity 0.000 traded against accuracy -1.030), and EFE is its on-policy extension, where the pragmatic drive -1.204 is reward-tilting and the epistemic drive 0.270 is active sampling over the student's own rollouts. As elsewhere, these are nats from a faithful minimal-model demonstration, not production measurements; the literature-reported empirics are reported separately.

9.0.1.1 Active rollout selection: expected free energy chooses where to distill

The energy ledgers above evaluate EFE on a fixed model; the *active* step is to let expected free energy choose which states the student rolls out on, which is precisely what the word on-policy names and what the correspondence map otherwise leaves to action selection rather than to the realized-rollout loss.

fig. 18 makes that selection explicit on a minimal T-maze-style menu of data-collection policies. A student that matches the privileged teacher posterior on a policy's own observation distribution keeps an expected residual uncertainty about the reward-relevant latent equal to that channel's conditional entropy, so the student–teacher gap a policy can close is exactly its epistemic value [Parr et al., 2022]: across the 6-point cue-validity sweep, epistemic value and residual gap sum to the prior entropy 0.693 nats at every point. Minimising expected free energy over the 4 candidate policies therefore selects the cue-visiting policy, whose informative rollouts drive the residual gap to 0.0e+00 nats; an ablated selector that keeps only the pragmatic reward-tilting term commits to an arm and leaves 0.693 nats of teacher signal unrecoverable, and blinding the cue reopens the gap. In this finite toy the epistemic term of expected free energy is thus not decoration but the quantity that fixes whether on-policy distillation can become exact; as elsewhere the claim is exact only for this declared model and is not a statement about production distillation.

The active-selection result is robust along three axes, each prototype-checked before promotion. First, **generality**: the identity is not an artefact of the binary toy – it holds across 6 observation channels with three- and four-valued latents to 5.6e-17 nats, while a wrong-measure ablation (weighting the residual by a uniform rather than the predicted observation distribution) breaks it by 0.019 nats, confirming the identity measures the genuine coupled quantity rather than holding vacuously. Second, **sequential depth**: a single step cannot show why an agent visits the cue *first*, since epistemic value is instrumental to a later goal. With a declared cue step-cost of 1.00 nats, a myopic one-step planner prefers to commit immediately while a two-step sophisticated-inference planner still prefers the cue (1.198 versus 1.722 nats of summed policy expected free energy) – the planning horizon is what creates the preference. Extended over a horizon, the cue's instrumental value scales: the cue/commit expected-free-energy gap grows by exactly 0.830 nats per remaining exploit step (break-even horizon 1.33), so a one-step planner commits but every horizon of two or more visits the cue, with the cost held in an analytically-derived window rather than tuned. Third, and most striking, the closed-form result **quantitatively predicts a measured observable of the project's pymdp simulation** (fig. 19): the analytical residual at the environment's own cue validity (0.95) equals the sophisticated-inference agent's measured post-cue belief entropy (0.199 nats) to 6.6e-09 nats, the agent visits the cue before any arm, and the match holds only at the environment's actual cue validity – a prediction bound to the model, not a fit. The prediction is not confined to that single post-cue point: the closed-form running-Bayesian belief entropy tracks the agent's measured belief entropy across the whole rollout – the prior, the post-cue plateau, and the resolved-to-zero tail – to 6.6e-09 nats (fig. 20), with the non-trivial agreement at the cue transition itself; a wrong cue validity or a shuffled observation order breaks the match. The bridge is stated at the level of observable behaviour and belief entropy (pymdp does not expose its internal expected-free-energy terms), and remains exact only for these declared finite models. The algebraic *structure* of the underlying complement identity is additionally machine-checked: a sorry-free Lean skeleton certifies $I(o;r) + E_o[H(r \mid o)] = H(r)$ over the integers, bracketed by a positive witness (a residual-zero cue transfers the full prior entropy) and a negative control (a strictly positive residual transfers strictly less), and bounded $0 \le I \le H(r)$ — the integer image of the epistemic-value bound the sweep exhibits (sec. 7); the real-valued entropy form remains the two-route numerical witness above.

Taken together these results form one auditable picture rather than a list, and fig. 21 collects the whole set in a single view. The passive half — the per-token reverse-KL distillation loss is variational free energy — and the active half — expected free energy chooses where the student rolls out, scales with the planning horizon, and predicts the sophisticated-inference agent's belief trajectory — are checked alongside the analytical mutual-information cross-check, the reverse-KL/free-energy convergence, and the multi-state generalisation. Across the 6 quantitative identities and cross-checks the largest residual is 3.6e-08 nats – each within a tier-aware tolerance, with the proved identities at machine zero and the numerical witnesses at optimizer/inference float noise – and each of the 5 results carries a negative control that bites by a measured margin of at least 0.019 nats, so no result is green-by-construction. (Results that report a direction or reduction rather than an exactness residual, such as the sequential-shift loss drop, are reported in their own sections and are not precision rows here.) This is the contribution the finite models are built to make: not a claim about production distillation, but a fully audited correspondence whose listed quantitative claims are exact in the declared models and whose every result is falsifiable by a control that fires.

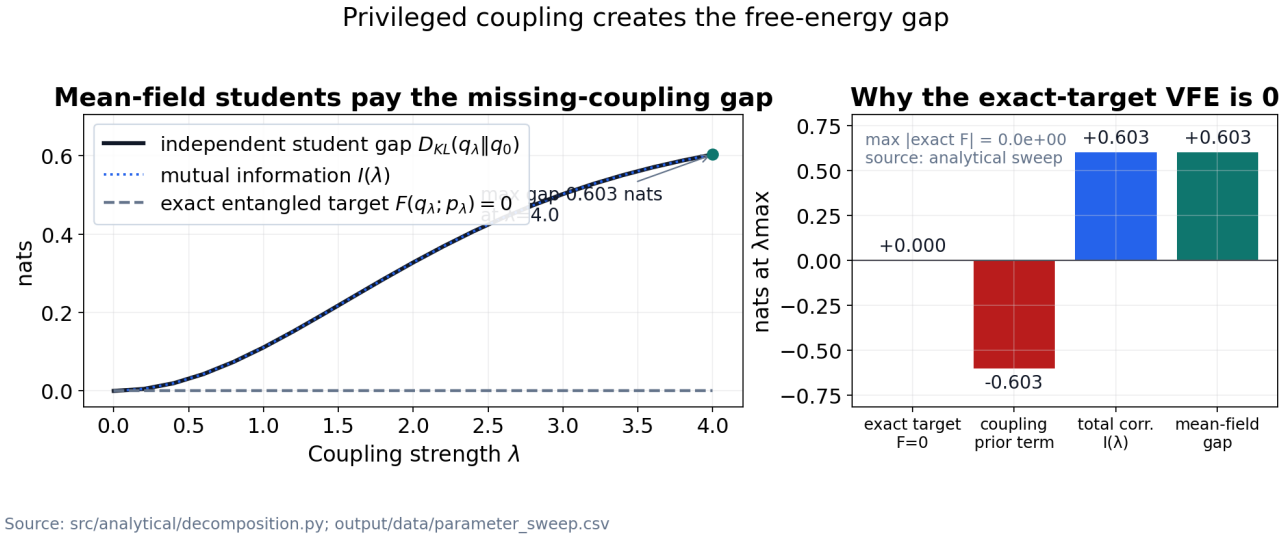


Figure 16: Free-energy gap created by privileged teacher–student coupling across 21 sweep points. The exact entangled target p_λ gives $F(q_\lambda; p_\lambda) = 0$ to numerical tolerance (max absolute value 0.0e+00 nats) because the posterior is evaluated against its own normalized target. The independent mean-field baseline q_0 instead pays $D_{\mathrm{KL}}(q_\lambda \| q_0)$, rising to 0.603 nats; in this symmetric toy that gap equals mutual information $I(\lambda)$ up to 4.4e-16 nats. The right panel makes the cancellation explicit: the exact-target coupling-prior term and total-correlation gain cancel, while the mean-field student keeps the information gap that on-policy distillation is meant to close.

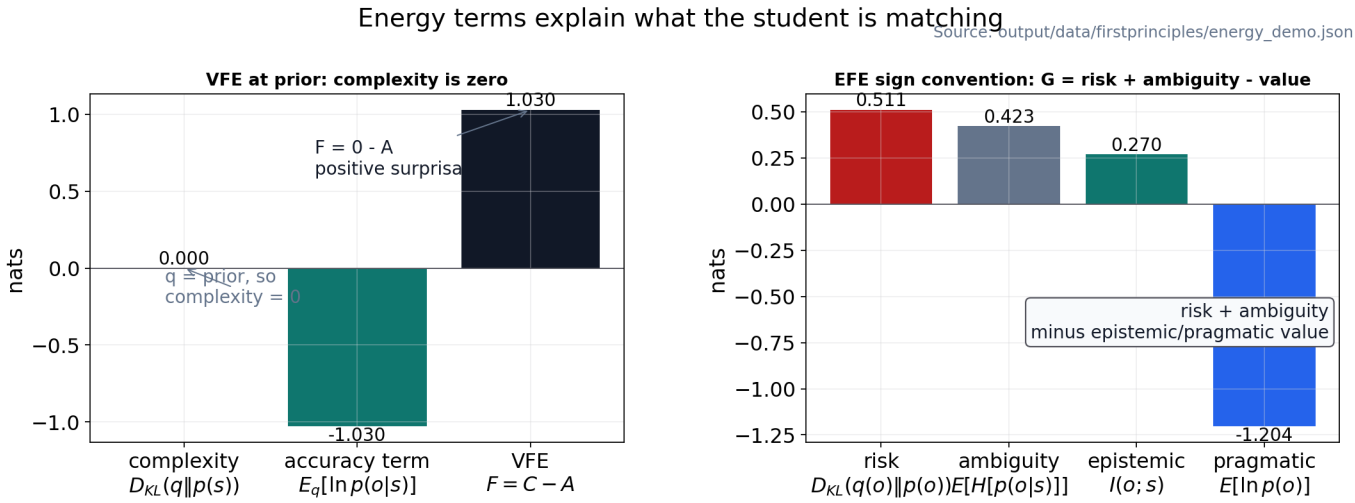
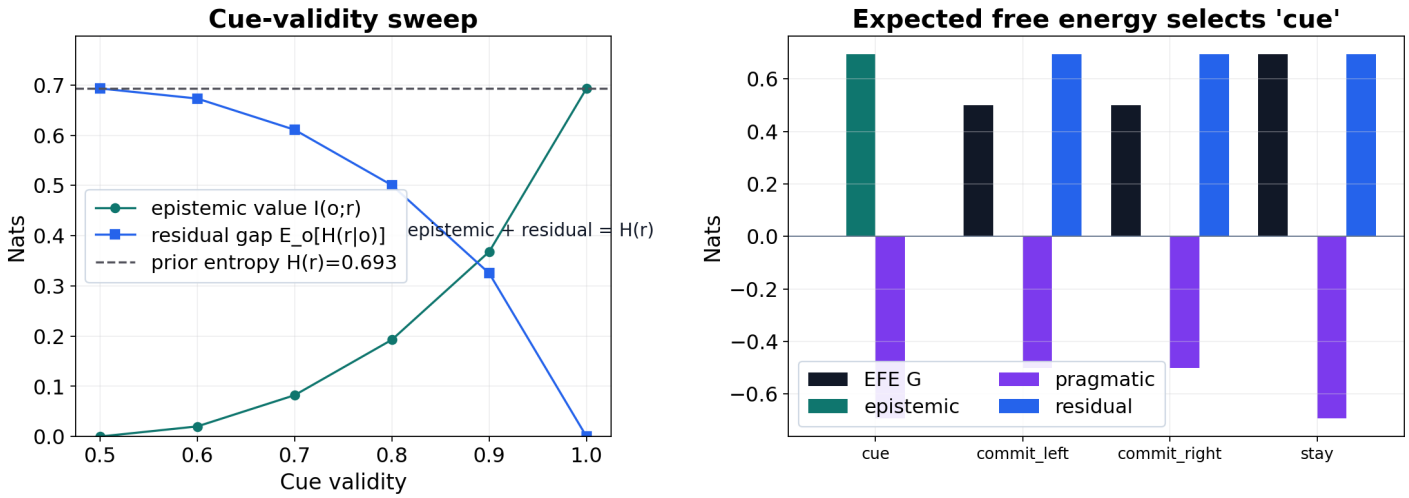
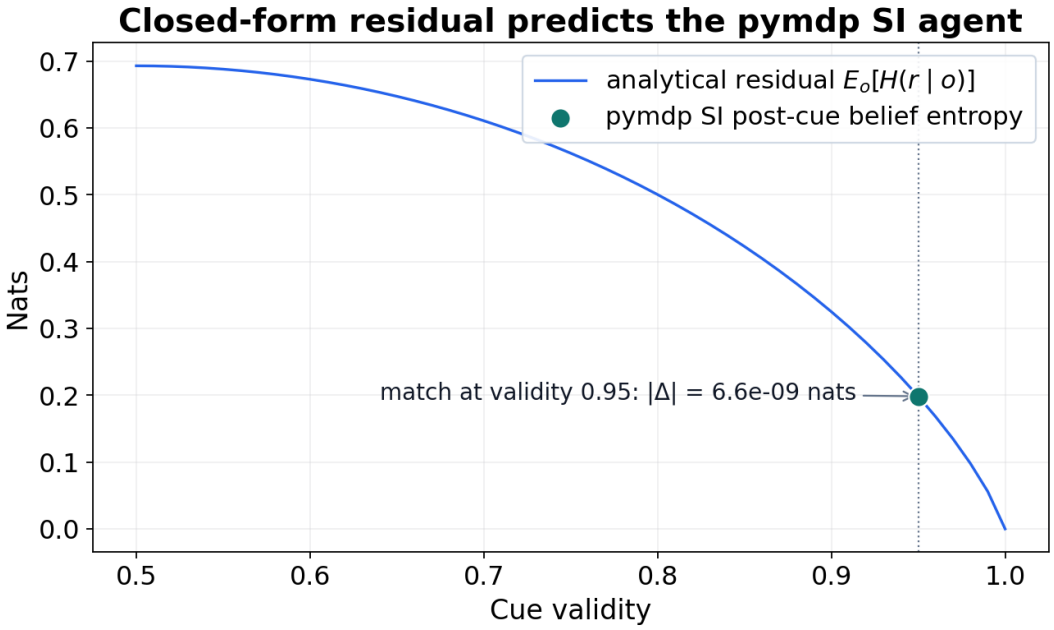


Figure 17: Energy-based decompositions for the categorical generative model, showing the two quantities active inference uses to score perception and action. Left: variational free energy split into complexity 0.000 minus the accuracy term -1.030 nats (bounding log-evidence -0.693). Complexity is zero here because the evaluated student belief equals the prior, so $D_{\mathrm{KL}}(q\|p(s)) = 0$; the VFE is therefore the negative of the negative log-likelihood accuracy term, not an unexplained zero. Right: expected free energy split into risk 0.511 plus ambiguity 0.423 nats, equivalently subtracting epistemic 0.270 and pragmatic -1.204 value. Sign key: risk and ambiguity are penalties (positive bars worsen G), epistemic and pragmatic are values (positive bars improve G), under the exact identity $G = risk + ambiguity = -(epistemic + pragmatic)$ as decomposed in `output/data/firstprinciples/energy_demo.json`; all bars are deterministic closed-form evaluations in nats. Reading distillation through this lens, the complexity/accuracy and risk/ambiguity terms make explicit that an on-policy student is simultaneously matching the teacher's outputs (accuracy/pragmatic) and reducing its own uncertainty about unvisited states (epistemic).



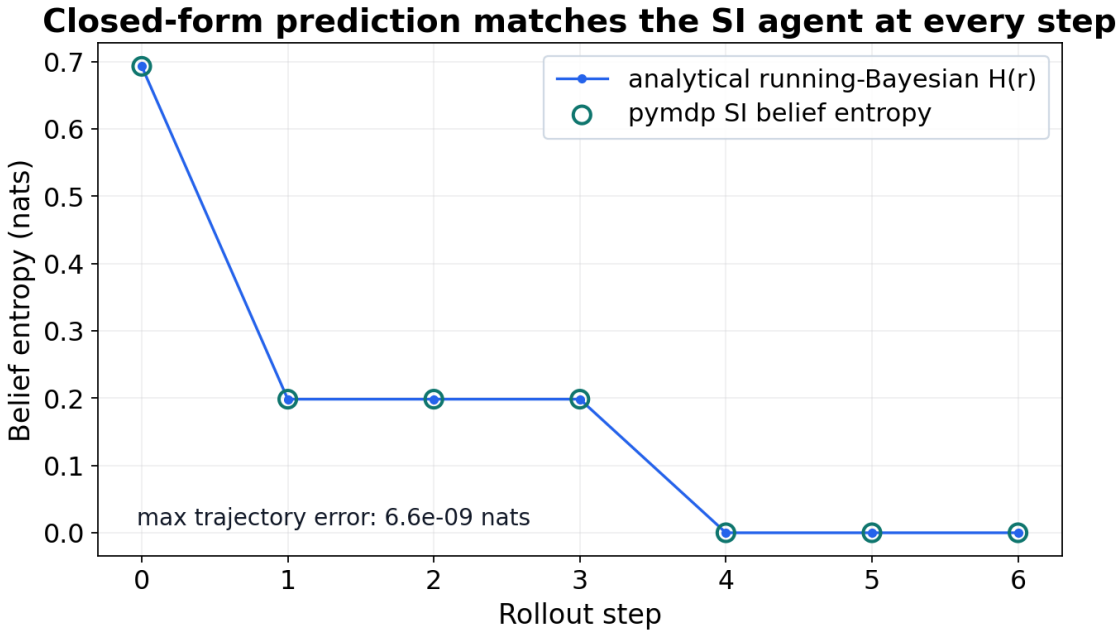
Source: output/data/firstprinciples/active_selection_demo.json; finite flat-prior toy, exact.

Figure 18: The active half of the correspondence: expected free energy chooses where the student collects rollouts. Left: across 6 cue-validity settings the epistemic value (information gain) rises and the residual distillation gap falls, and at every point their sum is the prior entropy H(r)=0.693 nats — the exact identity that the gap a policy can close equals its epistemic value. Right: the expected-free-energy decomposition for the 4 canonical data-collection policies; minimising expected free energy selects the cue, whose rollouts close the gap to 0.0e+00 nats, while a pragmatic-only rule commits to an arm and leaves 0.693 nats unresolved. Finite flat-prior toy, exact. Source: `output/data/firstprinciples/active_selection_demo.json`.



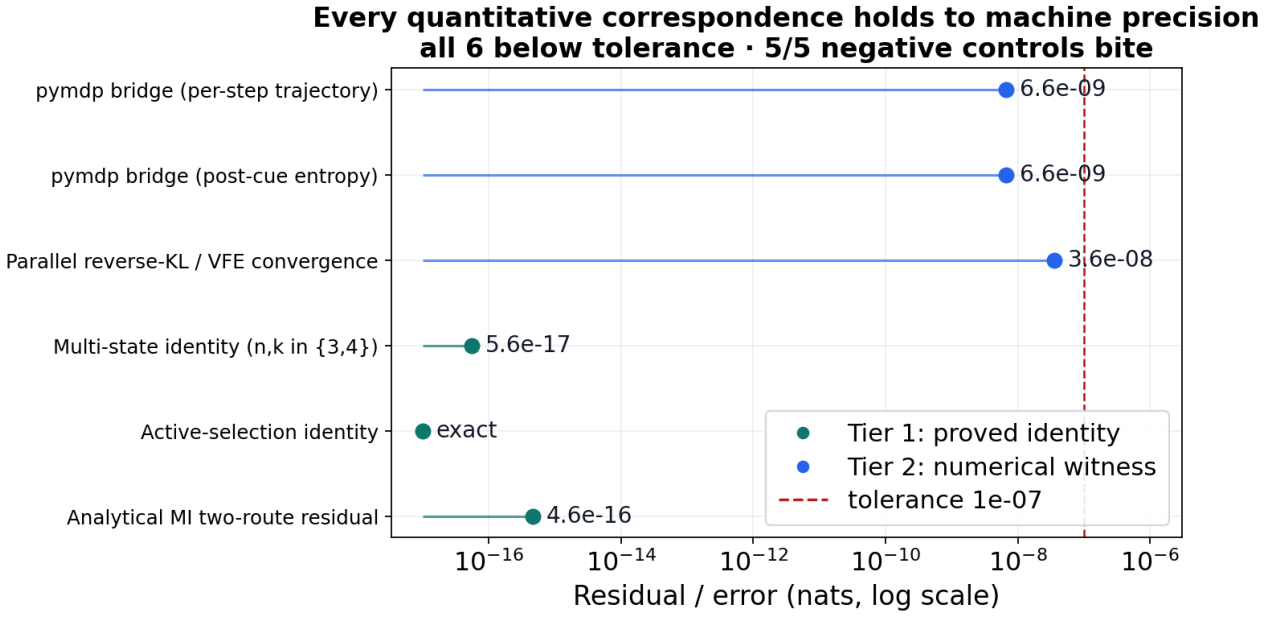
Source: output/data/firstprinciples/{si_bridge_demo,active_selection_demo}.json; finite toy, observable bridge.

Figure 19: The analytical active-selection result predicts the project's pymdp sophisticated-inference simulation quantitatively. The curve is the closed-form residual $E_o[H(r \mid o)]$ versus cue validity; the point is the SI agent's measured post-cue belief entropy (0.199 nats) at the environment's own cue validity (0.95). They agree to 6.6e-09 nats — a quantitative, validity-specific prediction, not a fit, and the agent visits the cue before any arm. The bridge is bound to observable belief entropy because pymdp does not expose its internal expected-free-energy terms. Finite toy, exact. Source: `output/data/firstprinciples/si_bridge_demo.json`.



Source: output/data/firstprinciples/si_bridge_demo.json; finite toy, observable bridge.

Figure 20: The closed-form bridge predicts the pymdp sophisticated-inference agent's belief entropy at every step, not only post-cue. The line is the analytical running-Bayesian belief entropy over the reward-location latent (flat prior, sharpened by the cue observation, resolved by the reward); the open circles are the agent's measured belief entropy. They agree across the whole rollout to 6.6e-09 nats, and a wrong cue validity or shuffled observation order breaks the match. Observable bridge, finite toy. Source: `output/data/firstprinciples/si_bridge_demo.json`.



Source: output/data/firstprinciples/precision_ledger_demo.json; synthesis over the result set.

Figure 21: A synthesis over the result set: every quantitative correspondence in the paper plotted by its residual or error against the tolerance line. All 6 of these quantitative identities and cross-checks – the analytical mutual-information cross-check, the active-selection and multi-state identities, the reverse-KL/free-energy convergence, and the post-cue and per-step pymdp bridges – fall below their tier-aware tolerance, with maximum residual 3.6e-08 nats, and each of the 5 negative controls bites by a measured margin. Tier 1 rows are proved closed-form identities; Tier 2 rows are numerical witnesses. Finite toys, exact in the declared models. Source: `output/data/firstprinciples/precision_ledger_demo.json`.

10 On-policy student rollout (T-maze)

The T-maze rollout instantiates one process-level witness for the thesis: an agent that generates its own observations and acts to minimise expected free energy, the active-inference analogue of a student policy producing the rollouts on which a teacher scores it [Agarwal et al., 2024, Friston et al., 2017a, Parr et al., 2022, van Oostrum et al., 2024]. Sophisticated inference here is only the pymdp planner's finite beliefs-about-policies machinery; it is cited alongside self-distillation methods that condition teachers on student traces or predictive signals, but it is not evidence about those systems [Zhao et al., 2026, Liu et al., 2026e,c]. The pymdp harness rolls out the full TMaze active-inference agent under the canonical `sophisticated_inference` planner with SI search horizon 5 and Agent policy length 1. Summary metrics land in `output/data/si_tmaze_summary.json`; trace-level q_π rows, action marginals, modality observations, and tree metadata land in `output/data/si_tmaze_trace.json`.

In this correspondence privilege is operationalized as *differential cue reliability*: the cue observation plays the role of the privileged information I, and `cue validity` sets how reliable each agent's access to it is. In the toy world the cue is part of the rollout dynamics for every agent — what differs is access quality, a structured-partial-observation analogue (rather than a literal train-only variable removed at deployment) of a hint, verified trace, long context, visual clue, or rich textual feedback in privileged-information and self-distillation settings [Kirchhoff et al., 2018, Vapnik and Vashist, 2009, Lopez-Paz et al., 2016, Cai et al., 2024, Snell et al., 2022, Ye et al., 2026, Lazaridis et al., 2026, Liu et al., 2026a, Hübotter et al., 2026, Shenfeld et al., 2026]. Acting to disambiguate the cue is the epistemic component of expected free energy: the on-policy student seeks the teacher signal precisely on the novel states it reaches itself [Friston et al., 2017b, Champion et al., 2024, Sajid et al., 2021b]. The epistemic term offers one formal lens — not the demonstrated causal mechanism — for understanding why self-generated rollouts can expose mismatch that teacher-forced evaluation may hide; it is the information-gain term that off-policy, teacher-generated data does not supply [Ross et al., 2011, Sun et al., 2017, Bengio et al., 2015, Arora et al., 2022, Rohatgi et al., 2025, Pozzi et al., 2025, Hinton et al., 2015, Friston, 2010, Millidge et al., 2021b]. Rollout transitions: 6; recorded timesteps: 7. Mean belief entropy: 0.1841 nats; mean policy entropy: 0.7165 nats. Belief entropy over the rollout is traced in fig. 22; modality-specific observations and selected actions are in fig. 23. The initial selected action is `move_to_cue`, with cue-directed marginal probability 0.545; the cue appears at recorded timestep 1, the reward/outcome appears at timestep 4, and the extracted trace therefore records cue-before-reward ordering as true. fig. 24 shows how q_π action probabilities evolve across the rollout, while the policy-entropy drop after the cue is 0.7091 nats.

The policy posterior itself — the q over policies that this process witness uses as the distillation-target analogue — is shown measured, step by step, in fig. 29, drawn from all 14 of 14 grid rows in `output/data/pymdp_policy_posterior_grid.json`. The contrast between the two planners is the point: the sophisticated-inference posterior concentrates onto a single action within the first steps because the agent's own rollout delivers the cue observation that sharpens it, while the comparison-only vanilla evaluator remains near-uniform over its policy set for the whole horizon — it never acts on what it could learn. In this finite rollout, posterior sharpening under self-generated observations is the process-level motif: the variational posterior narrows because it generates the observations it is corrected against.

This single-agent rollout has a direct two-agent reading that we make explicit in the executable classroom demonstration: a privileged teacher with cue validity 0.98 against an on-policy student with cue validity 0.5 yields teacher belief entropy 0.247 nats versus the student's 0.347 nats. The measured effect is a toy posterior-sharpness gap induced by the teacher's stronger cue channel, not a prediction that a Markov blanket numerically fixes entropy in general. The advantage measured and claimed is posterior sharpness, not task success: in this 4-decision rollout the artifact records teacher goal-reached `false` and student goal-reached `true` - the sharper privileged posterior did not translate into goal attainment on this short horizon, and the manuscript claims only what the entropy series measures. The mean reverse-KL distillation signal between the two is 6.28 nats - the finite toy objective that maps onto variational free energy $F = D_{\mathrm{KL}}(q \,\|\, p(s \mid o)) - \log p(o)$ (KL target the exact posterior $p(s \mid o) \propto p(o, s)$) [Gu et al., 2024, Levine, 2018]. This multi-step classroom divergence is a per-rollout quantity on a different scale from the single-decision mutual-information ceiling of $\log 2$ nats in the Bernoulli toy (sec. 5); the two measure different objects — a trajectory-level distillation signal versus a per-decision information bound — and are not to be compared numerically. Entropy-aware and adaptive-exposure OPD work is therefore a design-space neighbour: it motivates why teacher uncertainty should sometimes switch a token toward mode-covering pressure or withhold unreliable supervision, while our classroom only reports the toy teacher/student entropy gap it generated [Jin et al., 2026, Han et al., 2026, Luo et al., 2026]. The classroom artifact is also where the manuscript keeps internal self-distillation separate from external privileged-information distillation: OISD-style internal alignment [Liu et al., 2026c] is cited as a method-family analogue, while this figure's numbers come only from `output/data/firstprinciples/classroom.json`. A teacher cue-validity sweep turns this single comparison into a dose-response experiment with its own built-in negative control (fig. 28). Across 6 privilege levels (student fixed at cue validity 0.5), the identical-agent baseline gap is 0.0 by construction — a wiring/fabrication check (identical configurations cannot differ), not a control for the effect itself. The belief-entropy advantage appears only at the top of the grid: the gap stays at zero through cue validity 0.9 and is +0.100 nats at 0.98. Whether the onset is a sharp threshold or a steep slope is resolution-limited here — the grid has a single level between 0.9 and 0.98 — so we claim only that the advantage is strongly nonlinear in cue validity. The mean reverse-KL distillation signal is the more sensitive detector of privilege, and the claim rests on its monotone rise across the sweep rather than any single level: above a 10^-3-nat floor (set four orders of magnitude above the about 10^-7-nat rollout float noise observed at low validities), the signal first clears the floor at cue validity 0.8 (0.0018 nats) and rises monotonically to 6.28 nats — on the thesis' own terms this is exactly what should happen, since the reverse-KL loss is the free-energy gradient the student would descend, and it registers privileged information that summary statistics of the posterior miss. As with everything in this section, these are deterministic toy measurements (gap/validity rank correlation 0.65), not significance claims.

The review-requested sequential-shift witness isolates the same train/test mismatch without sampling or production claims. In `firstprinciples.sequential_shift.v1`, a four-state/two-action student induces a test visitation distribution that differs from teacher-forced train visitation by shift mass 0.229. Evaluating the pre-correction student under teacher-forced visitation gives 0.333 nats,

underestimating its own induced test loss 0.409 nats; the deterministic on-policy correction reduces test loss to 0.096 nats, closing 0.313 nats (fig. 26). The companion `firstprinciples.sequential_shift_sensitivity.v1` sweep makes this less brittle by varying the correction fraction over 5 finite policy mixtures: induced test loss decreases from 0.409 to 0.096 nats and train/test shift mass decreases from 0.229 to 0.110 while every row remains normalized (fig. 27). This is a finite sequential-distribution-shift witness aligned with dataset aggregation / induced-distribution shift, covariate shift, privileged-information, and self-generated-rationale context [Ross et al., 2011, Shimodaira, 2000, Sharoni and Sabato, 2023, Zelikman et al., 2022], not an empirical OPD benchmark.

The matrix/value audit (fig. 12) confirms A=[[5, 5], [3, 5, 2], [3, 5, 2]]; B=[[5, 5, 5], [2, 2, 1]], normalized A/B/D probability mass, cue validity 0.95, and reward condition 0. Policy-comparison rows: 2 across canonical SI and vanilla comparison-only planners; goal-reaching rows: 1. Graph-world extension rows: 4 over 4 nodes, with goal-reached flag 1.

Beyond the static T-maze, a family of dynamic and finite sequential simulators stresses each leg of the correspondence under controlled distribution shift. Their convergent behaviour is a consistency check for the declared toy objectives, not a guarantee about stochastic OPD training. The first simulator isolates the *active-sampling* leg: a generalised-knowledge-distillation (GKD) sweep that scores the same teacher signal on student-generated versus teacher-generated rollouts [Agarwal et al., 2024]. Scoring the same student under its own state-visitation measure reveals 0.120 nats of loss, of which a teacher-visitation evaluation sees only 0.114 nats — an exposure gap of 0.006 nats that off-policy evaluation hides. This is the simulator's local information-gain analogue; it is not an empirical estimate of exposure-bias severity in language models [Ross et al., 2011, Sun et al., 2017, Bengio et al., 2015, Arora et al., 2022, Rohatgi et al., 2025, Pozzi et al., 2025, Millidge et al., 2021b]. The gap echoes the cue-disambiguation dynamics traced in fig. 22: the student must sample to be scored where it actually goes.

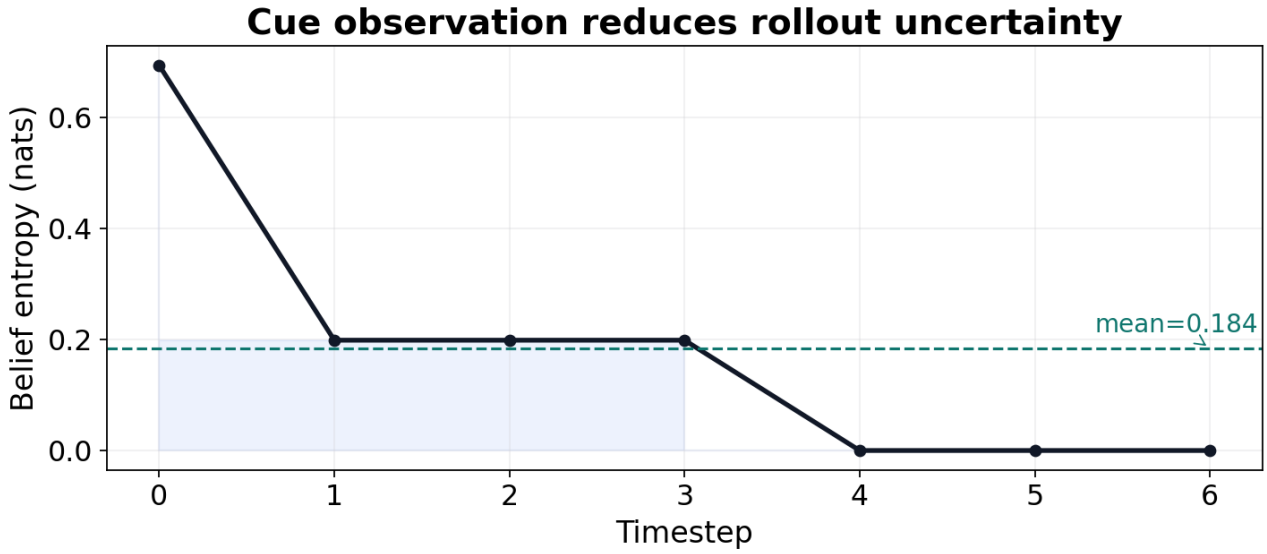
The second simulator isolates the *free-energy-descent* leg. Treating reverse-KL distillation as variational expectation-maximisation - E-step over the student's own rollout posterior, M-step on the generative-model parameters - the loop converges in 2 steps to a residual variational gap of 0.000 nats, and the descent is monotone (true), exactly the non-increasing free-energy trajectory expected under this clean E/M alternation [Friston, 2010, Friston et al., 2017a, Da Costa et al., 2020, Levine, 2018, Fellows et al., 2019]. The falsifiable signature here is local and algorithmic: this exact toy E/M objective should not increase across its own updates. It is not a claim that stochastic OPD training is globally monotone under arbitrary optimizers, teacher targets, or rollout distributions.

The third simulator isolates the *risk-tilting* leg through a diversity Pass-at-k sweep, in which the temperature of the student's sampling distribution trades finite-sample sharpness against coverage - the EFE risk/ambiguity balance read off generation diversity. For independent samples, $Pass@k = 1 - (1 - p)^k$, so the curve reports how the probability of at least one correct sample changes as the student distribution flattens or sharpens. The flattest (high-entropy, coverage-favouring) student reaches Pass-at-k 0.992 (fig. 37) versus the sharpest (low-entropy, concentrated) student at 0.865, while greedy decoding bottoms out at Pass-at-1 0.333; the ordering shows why a pure concentration objective can need an epistemic/coverage term in finite tasks [Gu et al., 2024, Hinton et al., 2015, Wu et al., 2024, Stanton et al., 2021, Jin et al., 2026]. Finally, an adaptive-divergence controller that interpolates between reverse- and forward-KL settles at a reverse-fraction of 0.50, an empirical compromise between concentration and mass-covering geometries reported separately in this manuscript and consistent with the variational/EFE blend rather than either extreme [Ko et al., 2024, 2025, Zhu et al., 2026b, Penaloza et al., 2026a, Parr et al., 2022]. Each simulator leg is a stylized minimal-model demonstration, not a production-scale measurement; their value is that they move in the direction the formal correspondence predicts inside declared finite artifacts.

We read these results as a minimal-model demonstration of the formal correspondence between on-policy distillation and active inference, not as claims about production language models; the quantitative findings are limited to the analytical toy, the T-maze rollout, the dynamic simulators, the sequential-shift witness and sensitivity sweep, and the classroom artifact reported here [Penaloza et al., 2026a, Da Costa et al., 2020].

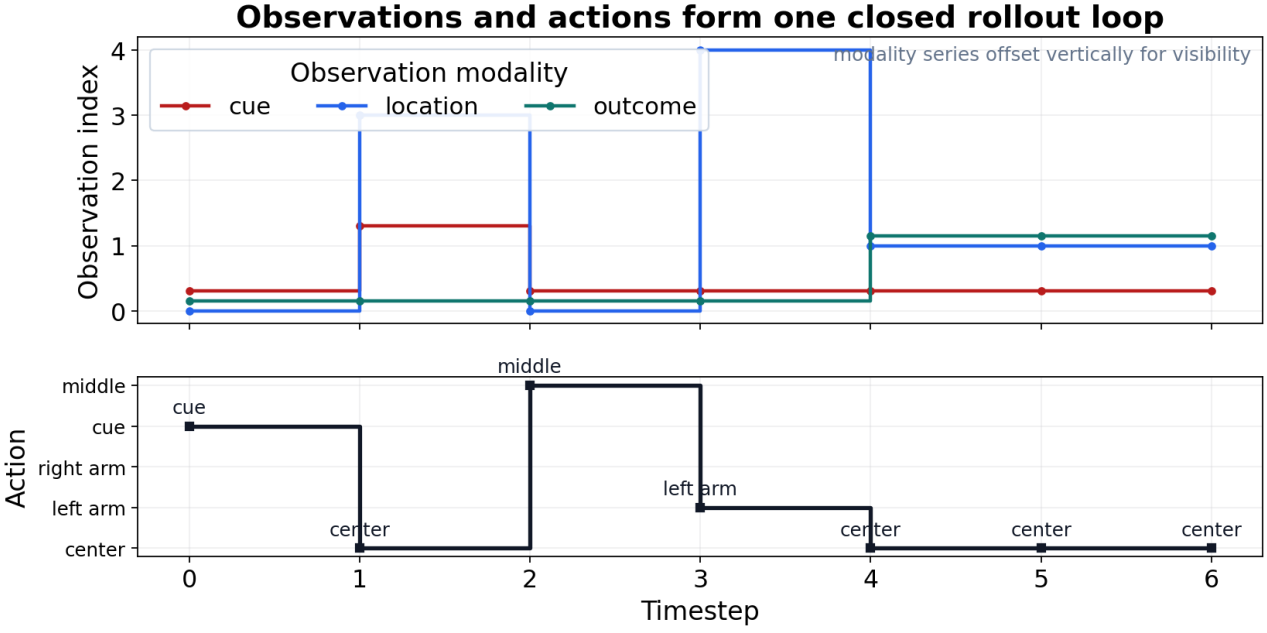
Rollout trace: `output/data/si_tmaze_trace.json`. Matrix/value audit: `output/data/si_tmaze_model_matrices.json`. JSONL run log: `output/logs/pymdp_runs.jsonl`.

See fig. 12 (Full TMaze generative-model matrix and value audit.)



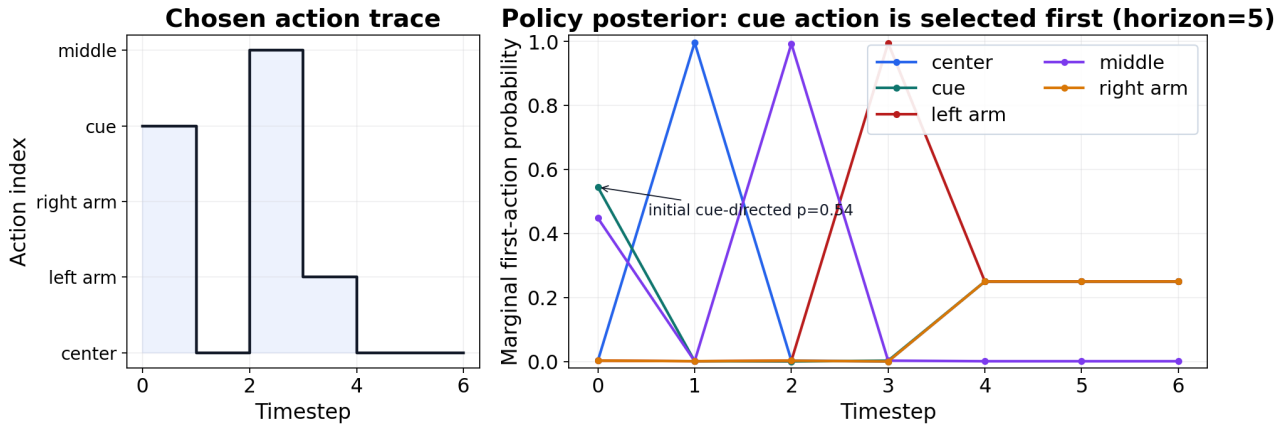
Source: output/data/si_tmaze_trace.json

Figure 22: Belief entropy in nats over the course of the pymdp T-maze rollout (mean 0.1841 nats; mean policy entropy 0.7165 nats), tracing how the agent's uncertainty changes once the cue is observed at timestep 1 and before the reward/outcome appears at timestep 4. The entropy trace is the epistemic payoff of the cue action made quantitative.



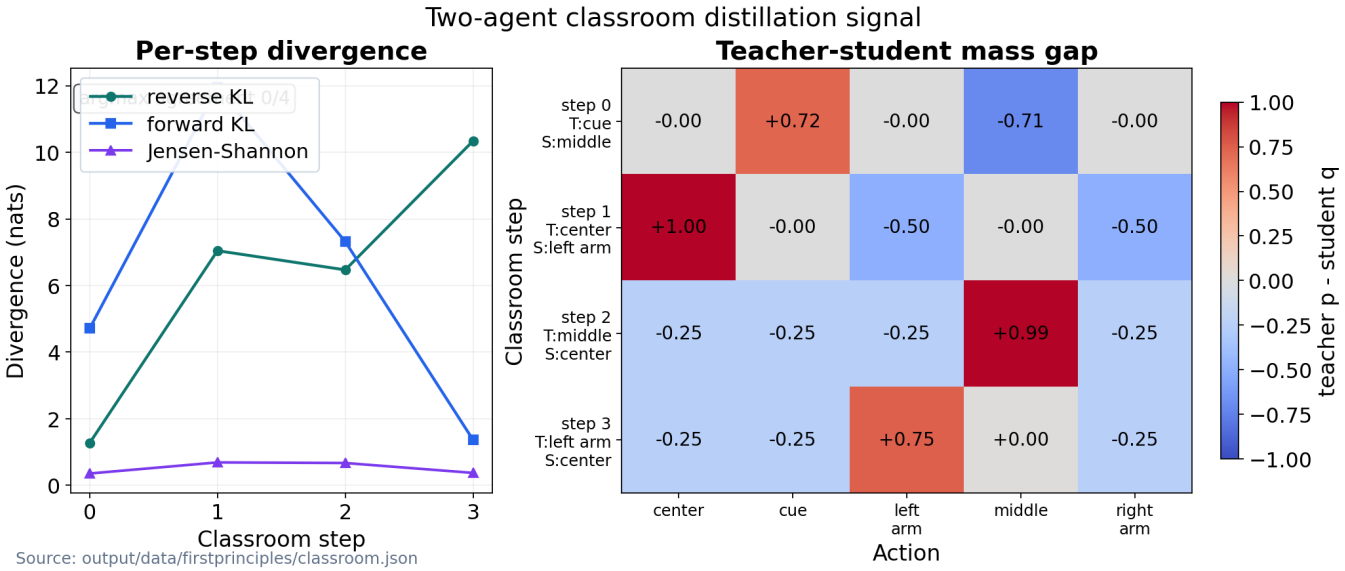
Source: output/data/si_tmaze_summary.json

Figure 23: Multimodal observation and action traces for the full TMaze rollout (3 observation modalities; action diversity 4). Upper: location, outcome, and cue observation indices over time. Lower: the selected action index and names. Together the panels show the closed perception–action loop that produces the teacher's behavior: cue observation at timestep 1, reward/outcome observation at timestep 4, and cue-before-reward ordering true. Capturing this joint observation–action structure, rather than a marginal action policy, is what makes the active-inference teacher faithfully reproducible.



Source: output/data/si_tmaze_summary.json

Figure 24: Canonical sophisticated-inference action selection for the full pymdp TMaze rollout (agent policy length 1, SI search horizon 5). Left: selected action index per timestep. Right: first-action marginals from the policy posterior over all five location actions. The initial selected action is move_to_cue, with cue-directed marginal probability 0.545. The policy-entropy drop after the cue is 0.7091 nats, quantifying the information-seeking step an on-policy student must learn to reproduce.



Source: output/data/firstprinciples/classroom.json

Figure 25: Two-agent finite toy classroom distillation signal over 4 decision points (3 transitions) between a privileged teacher (cue validity 0.98) and an on-policy student (cue validity 0.5). Left: per-step reverse KL, forward KL, and Jensen-Shannon divergence between the teacher and student policies (mean reverse KL 6.28 nats, mean Jensen-Shannon 0.522 nats, 0 agreement rows). Right: a heatmap of teacher-minus-student action probability by action and step, localizing where the student diverges. The figure operationalizes on-policy distillation as a per-step divergence signal evaluated along the student's own trajectory, the same active-inference principle of scoring beliefs on visited states rather than the teacher's idealized path. It is a deterministic toy signal, not an empirical OPD benchmark.

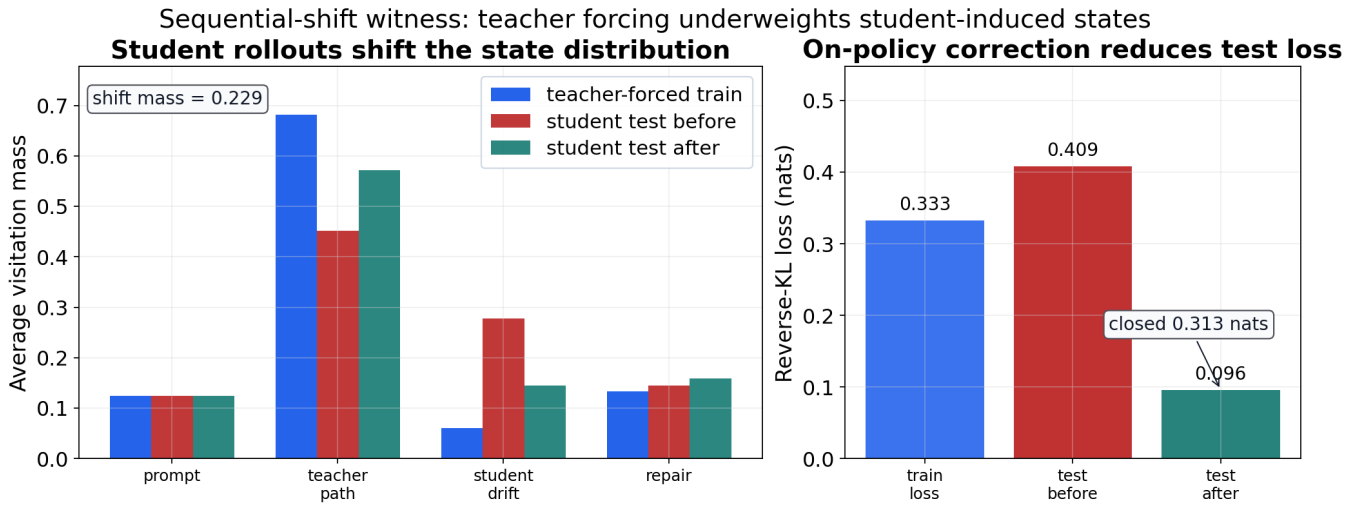


Figure 26: Deterministic finite sequential-shift witness requested by the critical review, not an empirical OPD benchmark. Left: teacher-forced training visitation underweights the student-induced `student_drift` state, creating shift mass 0.229 between the teacher-forced train distribution and the student's pre-correction test distribution. Right: evaluating the pre-correction student on teacher-forced states gives train loss 0.333 nats, which underestimates its own induced test loss 0.409 nats; the deterministic on-policy correction reduces the test loss to 0.096 nats, closing 0.313 nats. All probabilities are normalized finite rows from `output/data/firstprinciples/sequential_shift.json`; the panel is a toy witness for train/test distribution-shift accounting, not local evidence about production LLMs.

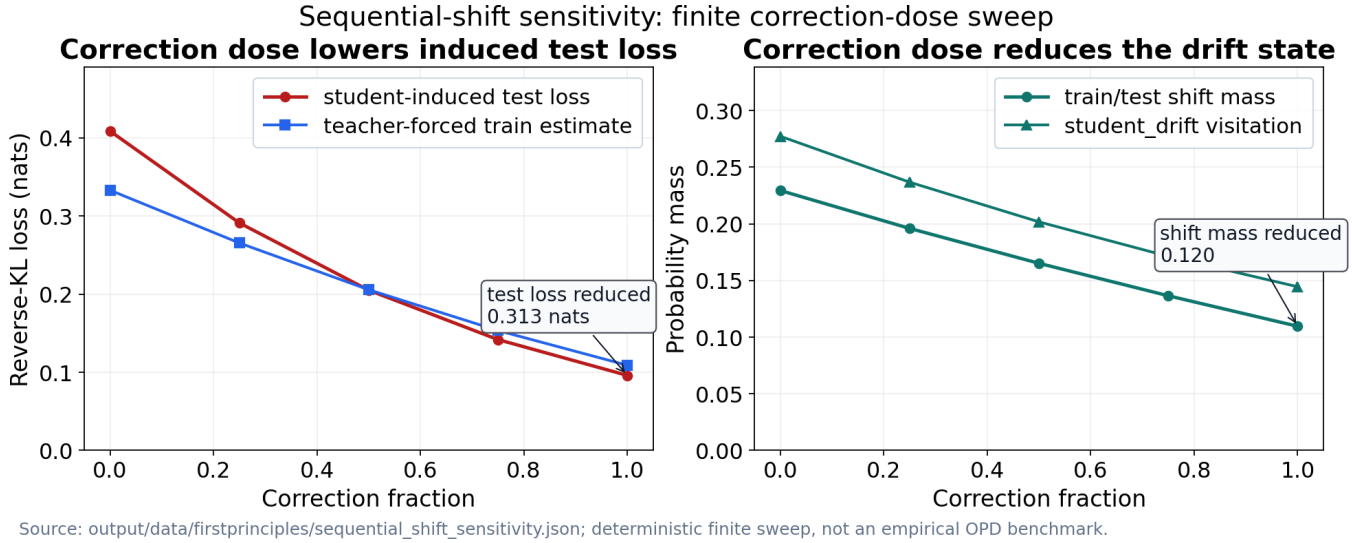


Figure 27: Deterministic finite correction-dose sensitivity sweep for the sequential-shift witness, not an empirical OPD benchmark. The sweep mixes the pre-correction and corrected student policies over 5 finite correction fractions, recomputes student-induced visitation at each fraction, and requires all policy and visitation rows to remain normalized. Student-induced test loss decreases from 0.409 to 0.096 nats, while train/test shift mass decreases from 0.229 to 0.110. Source: `output/data/firstprinciples/sequential_shift_sensitivity.json`; this is a sensitivity guard for the toy witness, not local evidence about production LLM optimization.

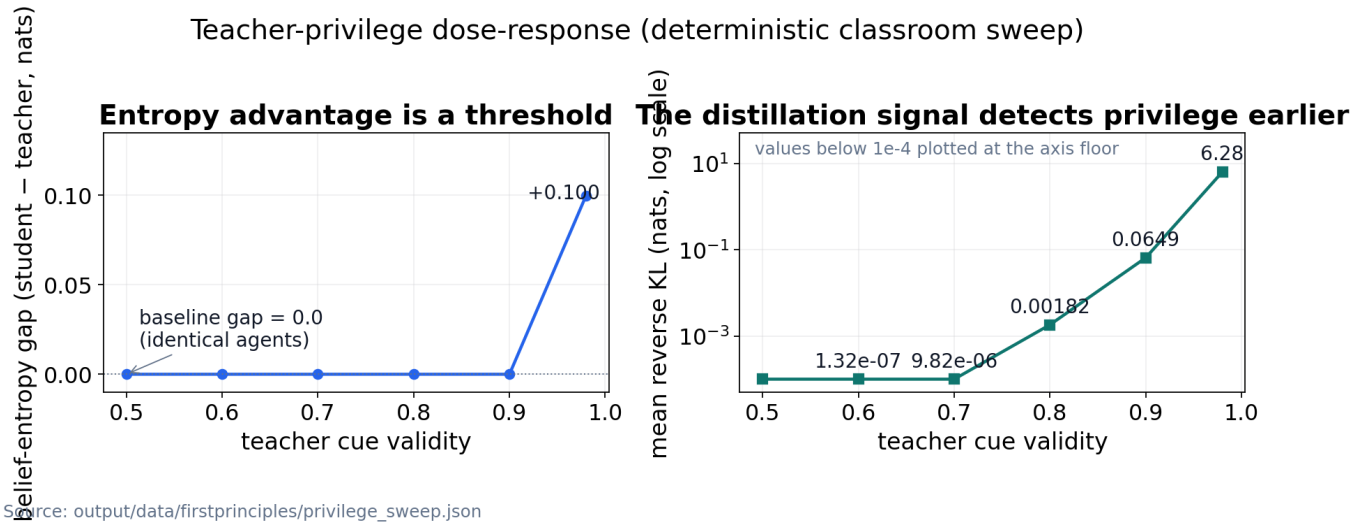


Figure 28: Teacher-privilege dose-response over 6 cue-validity levels (student fixed at 0.5; the identical-agent baseline gap 0.0 is a wiring check, not an effect control). The belief-entropy advantage is strongly nonlinear: zero through cue validity 0.9, +0.100 nats at 0.98 (step-versus-slope is resolution-limited by the grid). The mean reverse-KL distillation signal is the more sensitive detector, rising from its first appreciable value (0.0018 nats above a 10^-3-nat noise floor) at cue validity 0.8 to 6.28 nats — the loss the student would descend registers privilege that posterior-entropy summaries miss. Gap/validity rank correlation 0.65. Source: `output/data/firstprinciples/privilege_sweep.json`.

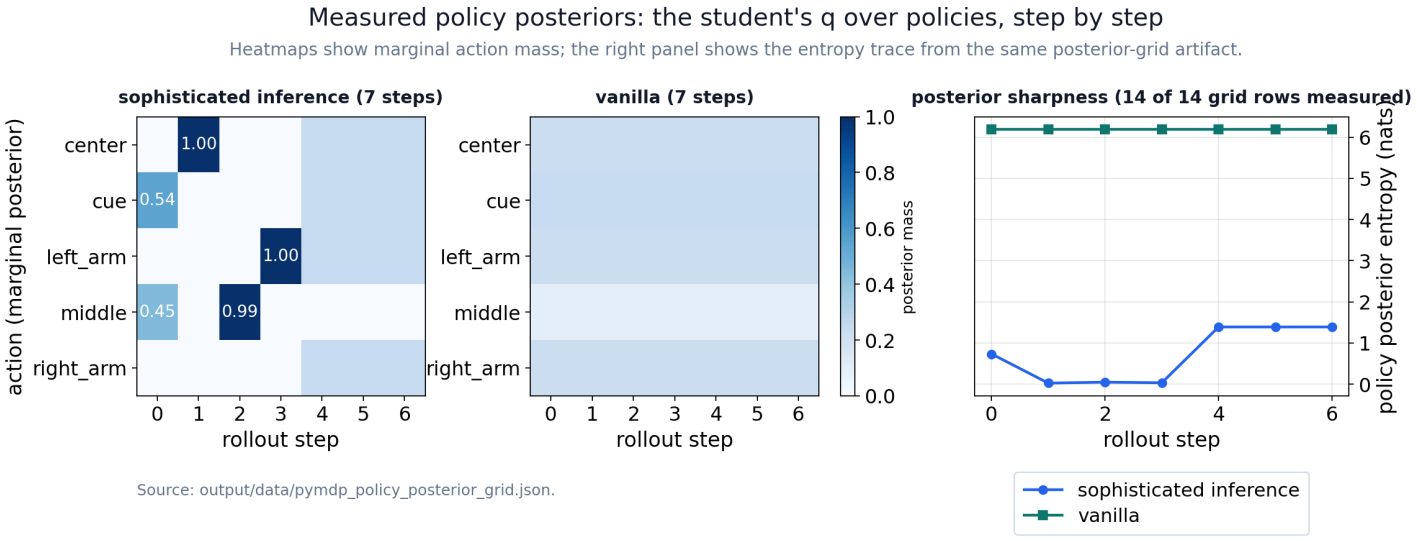


Figure 29: The student's measured posterior over policies, step by step — the quantity the correspondence identifies with the distillation target. Left and center: marginal action posteriors per rollout step for both planners (cell values are probabilities; labels shown for mass at or above 0.30). Right: policy-posterior entropy by step for all 14 of 14 measured grid rows. Entropy collapses after the cue observation: the on-policy rollout is what exposes the agent to the observation that sharpens its own posterior, the T-maze version of the induced-distribution argument for on-policy distillation. Source: `output/data/pymdp_policy_posterior_grid.json`.

Discussion

11 Limitations and outlook

11.1 What this supports

The result of this manuscript is a finite-model *correspondence*, supported under a verifier discipline, not a broad domain claim. In the finite objects studied here, the teacher policy plays the role of the intractable generative model, the student policy plays the role of the variational posterior, and per-token reverse-KL distillation on the student's own rollouts is variational-free-energy minimization by active sampling [Friston et al., 2006, 2009, Friston, 2010, Friston et al., 2017a, Sajid et al., 2021a, Parr et al., 2022]. The Bernoulli-Ising oracle supplies the analytical teacher-student coupling and mutual-information/free-energy identities. The pymdp T-maze supplies the on-policy active-inference rollout. The classroom toy supplies a two-agent teacher/student distillation signal. The sequential-shift witness and sensitivity sweep supply the review-requested finite train/test visitation mismatch check. The graph-world artifacts supply finite topology stress tests and Lean/model-checking witnesses. No gridworld result is part of the evidence surface. Across those scoped surfaces, every reported number is hydrated from a generated artifact, 6 sheaf axioms are machine-checked before composition, and 23 negative controls keep key failure paths live.

The divergence map is the practical design takeaway. In the finite examples here, the reverse-KL side concentrates on teacher-supported mass [Agarwal et al., 2024, Gu et al., 2024], the forward-KL side covers teacher-data mass [Hinton et al., 2015], and skew, entropy-aware, contrastive, or hybrid objectives occupy intermediate design points [Ko et al., 2024, 2025, Jin et al., 2026, Zhu et al., 2026b, Wu et al., 2024]. The alpha/f-divergence and KL-geometry literature is the necessary caveat: support, teacher entropy, optimization, clipping, and rollout distribution can change what the objective does in practice [Hernández-Lobato et al., 2016, Ke et al., 2019, Li et al., 2026, Luo et al., 2026]. Read this as divergence-geometry intuition rather than a universal mode-seeking law: which of mode-seeking or mass-covering behaviour a forward- or reverse-KL objective actually exhibits in language-model distillation is support-, student-expressivity-, and optimization-budget-dependent, so the finite map fixes the variational roles without fixing the empirical outcome [Wu et al., 2024, Gu et al., 2024]. The manuscript's contribution is not that these methods are empirically equivalent at scale, but that the finite artifacts make their shared variational roles explicit and auditable.

The same two design axes organise the literature itself. fig. 30 places all 37 methods of the audited taxonomy by publication year and by the on-policy/privilege quadrant each occupies: of the 37, 28 are on-policy and 13 condition the teacher on privileged signal, and the recent concentration in the joint quadrant — students generating their own training distribution under privileged teachers — is the regime the correspondence describes as a variational posterior generating its own observations under a generative model conditioned on privileged beliefs. The landscape is read directly from `output/data/firstprinciples/opd_taxonomy.json`; it is a positioning of the field's design choices, not a performance comparison.

11.2 Limitations

The Bernoulli-Ising toy, full TMaze harness, classroom run, graph-world extension, and sheaf composition model are pedagogical. They validate analytical consistency, artifact wiring, renderer dispatch, and manuscript hydration. They do not measure biological agents, cortical circuits, hierarchical transformers, production-scale distillation, or gridworld performance. The free-energy framing imports assumptions about a separable generative model, variational family, and observation boundary; the mathematical walkthrough literature is useful precisely because it makes those assumptions and limitations explicit [Millidge et al., 2021a]. The Markov-blanket and predictive-coding readings are likewise scoped to conditional-independence and top-down-target/bottom-up-residual roles — and because blanket definitions and free-energy derivations are not interchangeable without additional assumptions — a point the technical critiques make against the foundational formulation — we use them only as a constrained probabilistic interpretation of the toy models, not as a portable free-energy-principle derivation [Friston, 2010, 2013, Kirchhoff et al., 2018, Rao and Ballard, 1999, Biehl et al., 2021, Aguilera et al., 2022]. The exposure-bias motivation carries the same qualification: its empirical severity is task-dependent and autoregressive models can exhibit meaningful self-recovery, so the mismatch framing here is motivational rather than a universal diagnosis [He et al., 2021, Huszár, 2015].

The Bernoulli-Ising model is a faithful but minimal realization of teacher-student coupling: the coupling parameter is the channel by which the teacher's privileged variable informs the answer, the mutual information is the teacher-student mutual information, and the entangled-posterior free energy is the distillation objective. It is still a two-variable system, not a sequence model exhibiting induced-state distribution shift, exposure bias, tokenization, long-horizon credit assignment, teacher-selection sensitivity, teacher-entropy sensitivity, context-window shortcut risk, asynchronous freshness/staleness drift, or TopK-gradient instability [Pomerleau, 1989, Ross and Bagnell, 2010, Ross et al., 2011, Bengio et al., 2015, Arora et al., 2022, Pozzi et al., 2025, Buciluǎ et al., 2006, Hinton et al., 2015, Kim and Rush, 2016, Rusu et al., 2016, Czarnecki et al., 2019, Snell et al., 2022, Ye et al., 2026, Lazaridis et al., 2026, Jin et al., 2026, Chen et al., 2026, Zhu et al., 2026a].

The sequential-shift artifact adds a four-state/two-action induced-visitation mismatch check, and the sensitivity artifact verifies the correction direction across finite policy mixtures, but both remain toy accounting witnesses. They have no tokenization, long-horizon credit assignment, teacher selection, teacher entropy, context-window shortcut, asynchronous freshness/staleness, or production optimizer dynamics [Shimodaira, 2000, Sharoni and Sabato, 2023, Zelikman et al., 2022]. Likewise, the canonical pymdp planner is `sophisticated_inference` with SI search horizon 5: this on-policy agent generates its own observations and acts to minimize expected free energy, with the cue observation standing in for privileged information available in training but not at inference. The cue-disambiguation result is an epistemic-foraging toy, not evidence that LLM students will discover useful hidden structure at scale [Friston et al., 2017b, Tschantz et al., 2020a, van Oostrum et al., 2024]. The policy-comparison artifact exposes vanilla rows only as `comparison_only`

validation evidence, without changing the canonical rollout (sec. 6).

11.3 Threats to validity

The limitations above sort into the four standard validity buckets. *Internal validity*: every reported number is hydrated from a generated artifact and re-derived by a validator before render, with negative controls keeping failure paths live, so the dominant risk is in what the toys mean rather than in how they are computed. *Construct validity* is the gravest threat — the finite objects may not instantiate what the active-inference and on-policy-distillation vocabularies denote at scale; we bound this with the scoped Proposition and the differential-cue-reliability framing rather than a literal LUPI or Markov-blanket claim, and we keep the interpretive readings tagged as Tier 3. *External validity*: nothing here measures production-scale distillation, sequence models, tokenization, or long-horizon credit assignment, so the literature-reported rows below are neighbouring context, never evidence for this manuscript's claims. *Reproducibility validity*: the manuscript composes only when its artifacts, Lean inventory, and validation gates pass, and the artifact bundle ships with the paper, so the provenance claim is mechanically checkable rather than rhetorical.

11.4 Empirical evidence (literature-reported)

The structural correspondence suggests a mechanism someone could test at scale; the published on-policy distillation rows are compatible with that hypothesis but do not validate it for this manuscript. We did **not** measure any of the following ourselves; the table values in this subsection are from Table 21 of the Qwen3 technical report [Qwen Team, 2025], titled "Comparison of reinforcement learning and on-policy distillation on Qwen3-8B," as relayed and discussed by Thinking Machines [Lu and Thinking Machines Lab, 2025]. They are reproduced here only as external context for the correspondence, in the same spirit as the limitations above. On the AIME-24 mathematical-reasoning benchmark, Qwen reports on-policy distillation at 74.4 percent accuracy versus 67.6 percent for reinforcement learning — a gain of 6.8 points — while consuming 1800 GPU-hours against 17920 GPU-hours for the RL baseline, a compute reduction of 10.0x. Thinking Machines separately reports a replication at 70 percent AIME-24 in about 150 steps and frames the method as 9-30x more efficient than its RL comparison [Lu and Thinking Machines Lab, 2025].

A reduced AIME-24 excerpt of these literature-reported values is tabulated in the supplement (sec. 13); the complete source table — including the off-policy-distillation row and the GPQA-Diamond column that the excerpt omits — is the generated artifact `output/data/firstprinciples/benchmark_table.md` shipped with the manuscript.

The active-inference reading offers one interpretation of why such a gap could appear. Reinforcement learning supplies the student with a single sparse scalar — a reward at the end of a rollout — which in the free-energy view is an impoverished pragmatic signal that must be back-propagated across the whole trajectory before it shapes any token. On-policy distillation instead supplies a dense per-token free-energy gradient: at every position the teacher's privileged posterior defines a local target distribution, so the reverse-KL distillation loss yields an informative gradient at each token of the student's own rollout rather than one scalar per episode. This is the active-sampling regime in which the variational posterior generates its own observations and is corrected token-by-token against the privileged generative model [Friston et al., 2017a,b]. RLHF/instruction-tuning and self-generated-rationale systems are cited only as external context for this design space, not as locally reproduced evidence [Ouyang et al., 2022, Zelikman et al., 2022]. The Qwen-reported higher AIME-24 accuracy and 10.0x lower GPU-hour cost are consistent with that interpretation, but this manuscript does not isolate the cause at production scale. The active-inference lens is not the only candidate explanation, and the toy adjudicates between none of them: the same efficiency gain could arise from denser token-level supervision regardless of any variational reading, from a better-shaped verifier or teacher signal, from more favourable optimization geometry, from teacher-quality or curriculum effects, or simply from closer alignment between the student's induced rollout distribution and the target gradient [Wu et al., 2024, Jin et al., 2026, Han et al., 2026, Chen et al., 2026]. We therefore present the dense-per-token-gradient account as an interpretive lens (Tier 3), not a claim of causal sufficiency for OPD's empirical gains.

11.5 Audit, evidence, and open problems

sec. 1 and sec. 13 make binding state auditable under strict compose validation, with the reproducibility contract spelled out in the standalone supplement (sec. 14). The two-agent classroom simulation in `src/firstprinciples` (sec. 10) turns the same mechanism into a measured teacher/student entropy gap and reverse-KL distillation signal, while `firstprinciples.sequential_shift.v1` adds the finite induced-visitation mismatch check. Those signals exemplify the per-token objective shared by OPSD [Zhao et al., 2026], SDPG [Liu et al., 2026e,d], entropy-aware OPD [Jin et al., 2026], and internal on-policy alignment [Liu et al., 2026c], while all measured claims stay inside the generated toy artifacts.

The remaining open problems are exactly those the recent on-policy distillation literature has begun to chart but that these minimal models cannot settle. One cluster concerns scaling laws relating distillation temperature, teacher-student mutual information, and sample budget [Qwen Team, 2025, Lu and Thinking Machines Lab, 2025, Liu, 2026, Song and Zheng, 2026, Shrivastava et al., 2023]. Another concerns the Pass-at-1-versus-diversity-collapse tension and teacher/loss sensitivity that reverse-KL concentration can sharpen relative to mass-covering objectives — a tension adjacent to the broader neural-generation and KL-regularized RL literature showing that objective and decoding choices alone can induce low-diversity or mode-collapse behavior [Holtzman et al., 2020; Agarwal et al., 2024; Wu et al., 2024; GX-Chen et al., 2025; Stanton et al., 2021; Jin et al., 2026; Liu et al., 2026b; Zhu et al., 2026a; and the survey literature Liu, 2026; Song and Zheng, 2026].

Further open problems include skew-KL, sequence-level KD, contrastive KD, and adaptive or speculative reuse of student-generated outputs [Kim and Rush, 2016, Ko et al., 2024, 2025, Zelikman et al., 2022, Xu et al., 2024]; context and privileged-information OPD where a teacher can provide either transferable information or deployment-unavailable shortcuts [Snell et al., 2022, Ye et al., 2026,

Lazaridis et al., 2026, Liu et al., 2026a, Sharoni and Sabato, 2023]; black-box, trust-region, veto, reliability-restricted, position-weighted teacher-token reliability, and freshness-aware asynchronous OPD when rollout or teacher-supervision distributions become stale under student rollouts, together with offline approximations whose efficiency depends on teacher consistency with the student's induced distribution [Oh et al., 2026, Xing et al., 2026, Jang et al., 2026, Ye et al., 2025, Han et al., 2026, Liu et al., 2026b, Chen et al., 2026, Wu et al., 2026]; and stepwise, long-context, agentic, and multimodal OPD where variational-EM, RL-as-inference, maximum-entropy policy learning, DPO/RLHF, and joint OPD/RL objectives would have to be tested against multi-step rollouts [Fellows et al., 2019; Penaloza et al., 2026a; Penaloza et al., 2026b; Todorov, 2008; Toussaint, 2009; Ouyang et al., 2022; and the survey literature Liu, 2026; Song and Zheng, 2026].

Future work here remains modest: richer graph-world rollouts, larger formal bridges between the q_π trace and Lean witnesses, and expanded Lean proofs beyond the boundary witnesses in sec. 7. One epistemic note on the citation surface itself: much of the on-policy distillation literature engaged here is recent arXiv preprint work that has not yet completed archival peer review, so the field's empirical regularities should be read as provisional; the bibliography's source-kind classification in `output/data/scholarship_source_matrix.json` keeps the preprint/archival distinction machine-readable.

11.6 Toward LLM and world-model training runs

The correspondence is finite and audited, but it was built to be a lens on training regimes it cannot itself execute. This roadmap states, from first principles, what the variational reading would predict for on-policy LLM post-training and for world-model learning, and what would falsify it. Nothing in this subsection is measured here; the systems named are cited as external context for a future program, in the same literature-reported spirit as the limitations above.

An on-policy LLM training run, reduced to fundamentals. Strip an on-policy distillation run to its irreducible parts and four remain: a privileged teacher that defines a target distribution at every token, a student that generates its own rollouts, a per-token divergence between the two on the student's induced state distribution, and a gradient delivered at each position. These are exactly the variational objects the finite models instantiate — the teacher as the generative model, the student as the variational posterior, the per-token reverse-KL on self-generated observations as variational free energy minimized by active sampling [Friston et al., 2006, 2017a, Parr et al., 2022], and the privileged teacher signal as conditioning across a Markov blanket the student cannot cross at inference. The classroom statistics and the two-framework convergence run (sec. 6) show the identity holding where it can be checked exactly; the on-policy distillation systems now reported at scale [Agarwal et al., 2024, Gu et al., 2024, Qwen Team, 2025, Lu and Thinking Machines Lab, 2025] are the place the same decomposition would be measured rather than derived. Those systems are neighbouring empirical context for where the decomposition would be measured, not validation of the identities the finite models settle, and no scaling claim follows from the toy witnesses alone.

A world-model training run, reduced to the same fundamentals. A world model is, definitionally, a generative model of observations and dynamics; in the active-inference reading, learning one from experience is free-energy minimization and planning or acting with it is expected-free-energy minimization — the precise mechanism the T-maze toy already runs (sec. 10). This is a re-description in the framework's vocabulary, not a claim that systems trained at scale optimize a variational free energy by name: each minimizes its own evidence or return objective, which the active-inference idiom then reads as a free energy. Read this way the dominant world-model families can be seen as coordinate instances of the manuscript's generative-model term: recurrent latent imagination [Ha and Schmidhuber, 2018, Hafner et al., 2023], planning with a learned model [Schrittwieser et al., 2020], joint-embedding latent prediction [LeCun, 2022, Assran et al., 2023], and learned interactive environments [Bruce et al., 2024] each parameterize a model of observations and latent state and optimize a prediction-or-evidence objective that, in the active-inference idiom, is a free energy. The active-inference scaling literature makes the bridge explicit [Tschantz et al., 2020a, van Oostrum et al., 2024]: a world model is the generative model an agent minimizes free energy against, so a policy trained inside it is a variational posterior by construction.

The unifying conjecture. On-policy distillation and world-model learning are the same variational object approached from two ends. Distillation fixes the generative model — the teacher — and learns the posterior, the student policy; world-model learning learns the generative model itself. Active inference minimizes a single free energy over both, which yields this roadmap's central forward-looking hypothesis: that LLM post-training and world-model training are two coordinate-descent halves of one objective, and that on-policy distillation against a teacher conditioned on a learned world model is the natural bridge between them. The control-as-inference, trajectory-inference, and variational-EM literatures supply the formal scaffolding such a joint objective would inherit [Todorov, 2008, Toussaint, 2009, Fellows et al., 2019, Levine, 2018, Penaloza et al., 2026a], and the self-distillation and self-generated-rationale waves already gesture at alternating the two halves in practice [Zelikman et al., 2022, Xu et al., 2024, Jin et al., 2026].

Which parts are hard truths and which are open. Only one of the three correspondences is a mathematical fact about the objective itself; the other two are the constructed reading the scoped Proposition labels interpretive, and we are careful not to over-promote them. The proved, closed-form identity (Tier 1) is that reverse-KL on the student's own rollouts is the variational free energy of the declared objects, up to the evidence constant. That planning with a learned model is expected-free-energy minimization is a property of the pymdp *witness construction* — demonstrated by the rollout, and explicitly distinct from the realized-rollout objective the identity concerns (Proposition (iv)) — not a closed-form fact about the distillation loss. That privileged teacher information is conditioning across a statistical boundary is the interpretive Markov-blanket reading (Tier 3), scoped to conditional independence and stated with no physical or biological boundary claim. So the finite correspondence settles one algebraic identity, exhibits a second relation as a numerical/constructive witness, and offers the boundary reading as interpretation — not three facts of the same kind. Everything about behaviour at scale is hypothesis: that the dense per-token free-energy gradient is what buys on-policy distillation its sample efficiency over the single sparse scalar of reinforcement learning is consistent with the literature-reported runs but is not isolated here; that the identity survives sequence-scale induced-distribution shift, exposure bias, and long-horizon credit assignment is supported only by the finite sequential-shift witness; and the coordinate-halves conjecture itself appears only as two separately

validated halves, never jointly trained. Tokenization, context-window shortcuts, asynchronous teacher freshness and staleness, and TopK-gradient instability sit outside the toy entirely [Snell et al., 2022, Chen et al., 2026, Liu, 2026, Song and Zheng, 2026], named here as the boundary the correspondence makes visible rather than one it crosses.

A falsifiable program. Four scoped experiments would test the reading without inheriting any of its claims. First, instrument an existing on-policy distillation run as a per-token free-energy ledger: measure the reverse-KL-as-free-energy decomposition into accuracy and complexity, the teacher-student mutual information, and the epistemic-versus-pragmatic split of expected free energy, and ask whether per-token free-energy-gradient magnitude predicts the efficiency gain over reinforcement learning — the scaled-up form of the divergence-geometry and classroom-statistics artifacts here. Second, treat a learned world model as the teacher's generative model and distill a privilege-ablated student from it: a teacher conditioned on the realized latent or outcome, distilled into a deployable student that lacks that access, is the T-maze cue generalized to learned dynamics, and tests whether the privileged signal carries transferable structure or only a deployment-unavailable shortcut [Snell et al., 2022, Sharoni and Sabato, 2023]. Third, run the joint schedule directly — alternate a world-or-teacher-model update, the generative-model half, with on-policy distillation of the student, the posterior half — and check whether convergence and the Pass-at-1-versus-diversity tradeoff track the divergence-geometry predictions, with reverse-KL concentration trading against mass-covering forward-KL exactly as the finite divergence map sets out [Agarwal et al., 2024, Gu et al., 2024, Jin et al., 2026]. Fourth, promote privilege-ablation into an oversight criterion: a distilled advantage that vanishes when the teacher's privilege is removed is the signature of a shortcut rather than a learned mechanism, the scaled analogue of the scope-boundary distinction the audit already enforces.

The honest fence. None of these experiments is run here, and the correspondence does not predict their outcomes; it predicts their *form*. Scaling laws relating distillation temperature, teacher-student mutual information, and sample budget; the precise diversity-collapse threshold; tokenization and credit-assignment effects; and production optimizer dynamics all require the real runs and lie outside what any finite model can settle. The contribution this roadmap claims is narrower and, we judge, more durable: a single variational vocabulary in which LLM distillation and world-model learning are the same minimization, an explicit account of which of its assertions are identities and which are conjectures, and an audit discipline — the per-token free-energy ledger and the privilege-ablation control — that a scaled study could carry forward unchanged.

The discussion ontology binds `coverage_semantics` to the audit matrix in sec. 1, `pedagogical_scope` to the non-empirical scope of the toy models, and `sophisticated_inference_planner` to the pymdp harness contract in sec. 6.

Measured pymdp rollout (`sophisticated_inference`, config hash `1a6d58795fa5e8da`): mean belief entropy 0.1841 nats over 6 transitions and 7 recorded timesteps; goal reached flag 1; action diversity 4; SI tree available 1.

Analytical sweep residual RMSE 2.122461e-16 nats (max residual 4.440892e-16). Coverage audit: 95 present / 95 bound / 0 missing cells on the IMRAD matrix.

The scholarship matrix is also a scope-control device. It separates conceptual lineage from measured evidence: cited sources explain why the toy models are relevant, while generated artifacts decide every numerical, figure, and gate claim. That split keeps the paper from converting background authority into an unsupported empirical result.

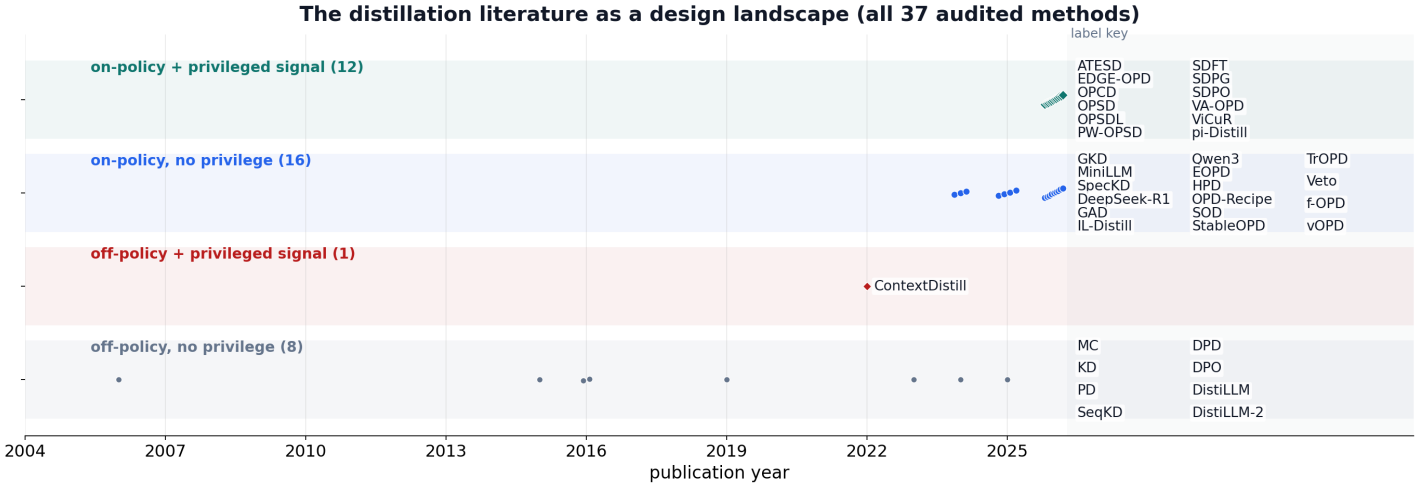


Figure 30: The distillation literature as a design landscape: all 37 audited taxonomy methods placed by publication year and by the two design axes the correspondence singles out — does the student generate its own training distribution (on-policy), and does the teacher condition on information the student cannot see (privilege)? Of the 37 methods, 28 are on-policy and 13 use privileged signal; the upper lane — both at once — is exactly the regime the active inference reading describes as a variational posterior generating its own observations under a generative model conditioned on privileged beliefs. Lane membership is read from the artifact, not hand-assigned; crowded lanes use a right-side label key, and remaining point labels use deterministic offsets, so horizontal position is approximate within one year. Source: `output/data/firstprinciples/opd_taxonomy.json`.

11.6.1 Ontology bindings

- `coverage_semantics` → **Coverage matrix semantics**

- `pedagogical_scope` → **Pedagogical scope**
- `sophisticated_inference_planner` → **Sophisticated inference planner**

11.6.2 Release notes evidence track

The `release_notes` track keeps release-language claims source-backed by validation, semantic, and bundle artifacts. Its evidence artifact is `output/reports/release_notes_evidence.json`: it currently records 3 rows, with source-backed status `true`.

12 Conclusion

This manuscript constructs and checks a finite-model active-inference reading of on-policy distillation. In the finite artifacts studied here, the teacher policy is read as the privileged generative model, the student policy as the tractable posterior, the reverse-KL distillation loss as variational free energy up to the evidence constant, and the student's own rollouts as the active samples on which that posterior is corrected. That is stronger than a slogan and narrower than a universal theorem. It is a claim about declared toy models, generated artifacts, and machine-checked manuscript bindings.

The Bernoulli-Ising oracle supplies the cleanest analytical witness. Its coupling λ is the teacher-student information channel, its mutual information $I(\lambda)$ is an interpretable ceiling on privileged information in this binary toy (no general communication bound is claimed), and its mean-field free-energy gap is the cost paid by a student that cannot condition on the teacher's privileged variable. The MI sweep appears once: a closed-form curve whose independent recomputation enters as a machine-precision residual rather than an overlapping line. The free-energy decomposition then shows why the same quantity is both an information gap and a distillation objective in this minimal model.

The pymdp T-maze supplies the active-inference process witness. Under canonical `sophisticated_inference` planning with config hash `1a6d58795fa5e8da`, the agent generates observations, samples a cue, and updates beliefs along its own visited trajectory. The cue is the toy privileged-information channel, not a production trace or hidden LLM feature. The classroom artifact turns that same mechanism into a two-agent distillation signal: a privileged teacher with cue validity 0.98 and an on-policy student with cue validity 0.5 produce a mean reverse-KL signal of 6.28 nats. The sequential-shift artifact closes a specific review gap without changing the scope: in a four-state/two-action finite witness, teacher-forced train visitation underweights the student-induced test states, train loss 0.333 nats underestimates induced test loss 0.409 nats, and the deterministic on-policy correction reduces test loss to 0.096 nats. The graph-world artifacts are finite topology stress tests and Lean/model-checking witnesses; they are not a gridworld benchmark, and no gridworld result is claimed.

The Lean, GNN, ontology, sheaf, and gate layers supply the publication witness. They do not prove a general theorem about all active-inference agents or all distillation algorithms. They prove something operationally valuable for this manuscript: symbols, generated numbers, figures, source-backed bibliography rows, theorem rows, and hydrated prose are forced through one artifact contract before the PDF exists. sec. 1 reports the binding state, sec. 14 records the reproducibility contract, and sec. 15 merges the analytical and simulation checks. A sweep RMSE of 2.1e-16 nats and 16 / 16 passed invariants summarize the internal consistency of the declared toy surface.

The external reasoning-distillation results remain external. Qwen's OPD-vs-RL values and Thinking Machines' relay/replication context help explain why the correspondence matters, but this manuscript does not reproduce those production-scale measurements. It also does not make a biological claim about Markov blankets, cortical predictive coding, or living systems. Its contribution is the disciplined bridge: it shows how privileged teacher feedback, reverse-KL distillation, expected-free-energy sampling, and artifact-level verification can be read through one variational ledger without letting toy results masquerade as empirical scale claims.

Each intended audience receives a distinct, separable contribution. For machine-learning readers, the manuscript offers a variational reinterpretation of on-policy distillation: a precise statement of which OPD design choices (rollout source, divergence direction, teacher conditioning) correspond to which variational roles, usable as a design vocabulary without adopting any further active-inference commitment. For active-inference readers, it offers an executable distillation analogue: a minimal, fully generated stack — analytical oracle, sophisticated-inference T-maze, two-agent classroom, and sequential-shift witness — in which the formalism's objects are instantiated by a contemporary training-method reading rather than by a biological metaphor. For reproducibility readers, it offers the artifact discipline itself: a sheaf-indexed compose contract under which no number, figure, or claim can enter the PDF without a generated, machine-checked witness. The three contributions stand or fall separately, and the preceding sections state which evidence supports each.

The final takeaway is therefore precise. Within the declared finite objects, on-policy distillation supports a stronger reading than a loose analogy to active inference: when the model is declared, the posterior family is fixed, and the student is scored on observations it generated, the same variational roles are doing the work. The manuscript's engineering contribution is to make that scoped claim auditable: every number and figure is produced upstream, every section is hydrated from the same evidence surface, and the gates fail closed when the manuscript drifts. Within the explicit finite models constructed here, the correspondence is strongly supported and unusually auditable; outside them, it remains a structured family resemblance whose limits the preceding sections state directly.

Appendix

13 Supplementary material: full coverage and concordance

This section is the **composability proof** for the manifest-indexed sheaf model that carries the on-policy-distillation-as-active-inference argument: all 22 appendix-bound fragment tracks render into one flat manuscript section without section-specific compose branches. It is intentionally a coverage and concordance supplement, not the operational reproducibility-methods section; the latter follows in sec. 14. Just as the thesis holds that one variational free-energy functional governs both the generative model (the teacher policy) and the approximate posterior (the student policy) [Friston et al., 2006, Friston, 2010, Agarwal et al., 2024], one registry governs every fragment type here. The registry defines 33 composable types, and this row binds every registered fragment slot, including the generated `layers` tables and optional `animation` fragment, alongside the live proof, simulation, formal, notation, validation-spine, integration, audit, finite-catalog, ablation, license, release-evidence, scholarship, assumption-index, delta, and staleness tracks. The heterogeneous fragments are the manuscript-level analogue of the correspondence's heterogeneous terms — the Bernoulli–Ising coupling toy, the pymdp sophisticated-inference rollout, and the two-agent classroom run [Parr et al., 2022] — each carrying a distinct piece of evidence under a single composition law. The sheaf language follows a finite local-to-global and composition-contract discipline [Curry, 2014, Speranzon et al., 2018, Robinson, 2014, 2017, Phillips, 2020, Fong and Spivak, 2019, Rosiak, 2022, Cox, 2026], not an unmeasured cohomological claim.

Supplemental concordance and metadata tables. To keep the main body prose-led, the large concordance and metadata tables are kept as separate, single-source markdown files rather than inlined: the active-inference ↔ on-policy-distillation correspondence map (`output/data/firstprinciples/correspondence_table.md`), the on-policy-distillation method taxonomy (`output/data/firstprinciples/taxonomy_table.md`), the literature-reported empirical benchmark (`output/data/firstprinciples/benchmark_table.md`), and the integrated notation/formalism supplement (`docs/reference/notation-supplement.md`). Each is regenerated from the `firstprinciples` package and the manuscript variables, so the supplemental tables never drift from the artifacts that produced them. The GNN ↔ ontology concordance and the sheaf coverage/scholarship matrices remain in their generated form under `output/data/`.

The proof is a publication-systems check (eq. 4). It demonstrates that heterogeneous fragments share one registry, manifest, renderer dispatch path, coverage matrix, and hydration boundary; it does not assert that every track carries equal scientific weight, nor that the analytical and T-maze demonstrations [Da Costa et al., 2020] license claims beyond these minimal models and artifacts. The machinery guarantees only that the structural mapping - variational free energy to reverse-KL distillation loss, active sampling to on-policy student rollouts [Gu et al., 2024], and privileged information to teacher-side conditioning across a statistical boundary - is rendered coherently across tracks, leaving the scientific weight of each correspondence to the sections that carry it.

13.0.1 Supplemental table: energy decomposition

The full variational- and expected-free-energy decomposition for the minimal model (referenced from sec. 9.0.1) is tabulated here. As elsewhere, these are nats from a faithful minimal-model demonstration, not production measurements.

Functional	Stream A	A (nats)	Stream B	B (nats)	Scalar (nats)
VFE (F)	complexity	0.000	accuracy	-1.030	log-evidence -0.693
EFE (risk/ambiguity)	risk	0.511	ambiguity	0.423	—
EFE (epistemic/pragmatic)	epistemic	0.270	pragmatic	-1.204	—

13.0.2 Supplemental table: empirical OPD-vs-RL benchmark (literature-reported)

The literature-reported AIME-24 benchmark (referenced from the discussion) is tabulated here. These are external empirical results from Table 21 of the Qwen3 technical report [Qwen Team, 2025], relayed and discussed by Thinking Machines [Lu and Thinking Machines Lab, 2025], not measured in this manuscript; only the toy-model statistics reported elsewhere here are hydrated from our own generated artifacts.

Table 1: AIME-24 accuracy and training cost for on-policy distillation versus reinforcement learning. The table cells are attributed directly to Table 21 of Qwen Team [2025]; Lu and Thinking Machines Lab [2025] relays those Qwen values and separately reports a 70 percent AIME-24 replication in about 150 steps with a 9-30x efficiency range. These are external empirical results, not measured in this manuscript; only the toy-model statistics reported elsewhere here are hydrated from our own generated artifacts.

Quantity (literature-reported)	On-policy distillation	Reinforcement learning
AIME-24 accuracy (percent)	74.4	67.6
Accuracy gain over RL (points)	6.8	—
Training cost (GPU-hours)	1800	17920
Compute reduction vs RL	10.0x	1.0x

For each track $t \in \mathcal{T}_{\text{Full}}$, the appendix row binds a fragment path $f(t)$ and the composer emits `<!-- sheaf-track:t -->` before the rendered body. Generated renderers such as `section_figures` and markdown renderers pass through the same `resolve_track_body()` dispatch, so the appendix exercises the common compose interface rather than a bespoke appendix path.

$$|\mathcal{T}_{\text{Full}}| = 22 \tag{4}$$

The fragment registry defines 33 composable track types. This appendix binds the generated `layers` report and optional `animation` fragment; the deterministic GIF artifact in `tracks.yaml` `extension_tracks` is produced by the core analysis DAG and remains separate from this fragment slot.

Because this appendix binds every registered appendix track, it is the maximal publication stalk of the coverage presheaf and exercises every publication renderer through the common `resolve_track_body()` dispatch. The same compose path is gated by the 6 sheaf laws verified in sec. 14 (6/6 satisfied): the appendix section glues to a unique output (separation), occupies the terminal position of the linear extension under its own `appendix` group row (poset and gluing), binds only well-typed fragments (typing), and owns every fragment path it references (compositionality). No count in this appendix is hand-written; all are injected from the registry-backed oracle.

Analytical sweep artifacts feed sec. 8 and sec. 15; simulation invariants merge after sec. 10. No additional path listing is required beyond those Results sections.

The appendix `assumption_index` row points to `output/data/analytical_assumption_index.json`. It binds 7 finite Bernoulli-Ising assumption rows to 7 equation identifiers and generated artifacts, with indexed status `true`.

The point is to make analytical signposting mechanical. If an equation is added without an assumption row, or if a row loses its evidence artifact, the index gate fails and the manuscript cannot present the equation as part of the validated finite toy proof surface.

pymdp harness summary: `output/data/si_tmaze_summary.json` (mean belief entropy, action trace, q_π rows, SI tree flag). Matrix/value audit: `output/data/si_tmaze_model_matrices.json` (A=[[5, 5], [3, 5, 2], [3, 5, 2]]; B=[[5, 5, 5], [2, 2, 1]]). Runtime diagnostics: `output/reports/pymdp_runtime_diagnostics.json` (known warnings 2, tree warnings 39, unexpected warnings 0). Policy posterior grid: `output/data/pymdp_policy_posterior_grid.json` (14 rows). Full log: `output/logs/pymdp_runs.jsonl`.

`sheaf-track:interop` binds `output/data/interop_roundtrip_report.json`, `output/data/gnn_roundtrip_report.json`, `output/reports/gnn_lint_report.json`, and ontology profile artifacts into the appendix proof row. The appendix claim is exactly 6 checks with lossless status `true`.

The appendix provenance fragment points to `output/data/artifact_provenance.json`, the canonical artifact that records required toy artifact hashes, producer scripts, source commit, deterministic seeds, config digests, and 5 bundle rows.

`replay_matrix.json` provides the appendix proof for deterministic replay: 14 producer replay/fingerprint rows with matched status `true`.

The appendix counterexample fragment points to `output/reports/counterexample_matrix.json`, the expected-failure matrix that keeps promoted validation gates falsifiable. It currently records 23 known-bad fixtures, and the hydrated pass flag is `1`, meaning those fixtures are expected to fail rather than sneak through a positive-control gate.

This row is the negative-control ledger for the sheaf. Each counterexample names a promoted track, target validation gate, mutation, and observed expected-failure status. A new live track without a counterexample row is therefore visibly incomplete in the track-improvement scope.

`sheaf-track:adversarial_audit` binds `output/reports/adversarial_audit.json`, `output/reports/scope_boundary_audit.json`, and claim-audit outputs. The appendix claim is exactly 23 expected-failure rows with documented status `true` and known-bad-passing count 0.

`evidence_field_index.json` provides the appendix proof for field-level claim evidence: 145 mapped fields with status `true`.

`release_bundle_manifest.json` provides the appendix proof for required deliverables: 39 artifacts with source-present status `true`.

`validation_gate_index.json` provides the appendix proof for gate ergonomics: 27 indexed gates.

13.0.3 Appendix track: artifact diffoscope

`artifact_diffoscope` binds `output/reports/artifact_diffoscope.json` into the full sheaf appendix. Rows: 81. All equal: `true`.

This diffoscope is deliberately narrow and reproducibility-facing. For each non-cyclic generated artifact, it compares the saved provenance digest to the live file digest at validation time. The validator re-derives equality from the rows, so a stale `all_equal: true` summary cannot hide one changed artifact.

The row count is not a decoration; it is the number of artifact fingerprints that survived cycle exclusion and therefore can be compared directly. This keeps the release bundle honest about mutable files while avoiding self-referential hashes for artifacts that necessarily include their own provenance.

13.0.4 Appendix track: artifact license

`artifact_license` binds `output/reports/artifact_license_audit.json` into the full sheaf appendix. Rows: 121. All safe: `true`.

The license audit classifies each generated or source-backed artifact under the public study's configured license boundary. It is intentionally conservative: generated local outputs and project-owned source files pass, while an artifact outside those public source kinds would need an explicit provenance and license row before it could support a manuscript claim.

This is also where the blocked empirical-adapter boundary matters. Private, restricted, or network-derived data are not smuggled in as evidence; they remain blocked until privacy, licensing, typed-claim, semantic, and negative-control gates are implemented in the same artifact path.

`sheaf-track:scholarship` binds `output/data/scholarship_source_matrix.json` into the appendix proof row. The appendix claim is exactly 127 connected source rows with connected status `true`; each row names a bibliography key, method role, manuscript section, registered track set, evidence artifact, and claim-boundary statement.

`sheaf-track:sensitivity` binds `output/data/sensitivity_sweep.json`, measured `output/data/si_policy_grid.json`, compatibility-named EFE values artifact `output/data/si_efe_terms.json`, `output/data/analytical_observable_sweep.json`, and graph-world topology artifacts including `output/data/si_graph_world_topology_traces.json`. The appendix claim is exactly 96 complete canonical grid cells.

`sheaf-track:uncertainty` binds `output/data/uncertainty_summary.json`. The appendix claim is exactly 21 normalized rows across 3 entropy bins with status `true`.

`sheaf-track:benchmark` binds `output/data/toy_benchmark_matrix.json`. The appendix claim is exactly 3 complete toy-model rows with status `true`.

Figure 31: Theorem traceability graph generated from 22 linked theorem rows and 397 proof-dependency edges; all 22 theorem rows are drawn with their finite-witness counts, and all theorem rows have resolved dependency edges: true. The graph exposes the deductive backbone of the formal track – which lemmas each distillation/active-inference theorem rests on and which finite models witness it. Fully resolved dependencies show that each declared finite theorem row has a registered proof-dependency chain rather than appearing as an isolated assertion; they do not close formal obligations outside this inventory.

13.0.5 Appendix track: state-space catalog

`state_space_catalog` binds `output/data/state_space_catalog.json` into the full sheaf appendix. Rows: 6. All finite: `true`.

The catalog is the finite-scope boundary for every toy claim in the study. Each row records a model id, state count, action count, policy count, source artifact, and finite flag; the validator recomputes that counts are positive and that every row remains finite. This prevents a manuscript sentence about exhaustive checking from silently drifting into an unbounded or empirical setting.

`output/data/state_transition_table.json` makes the boundary operational. It contains 24 deterministic transition rows and covers all reachable finite models with status `true`. Readers can therefore audit not just the number of states, but the actual state/action/next-state relation used by the model-checking witnesses.

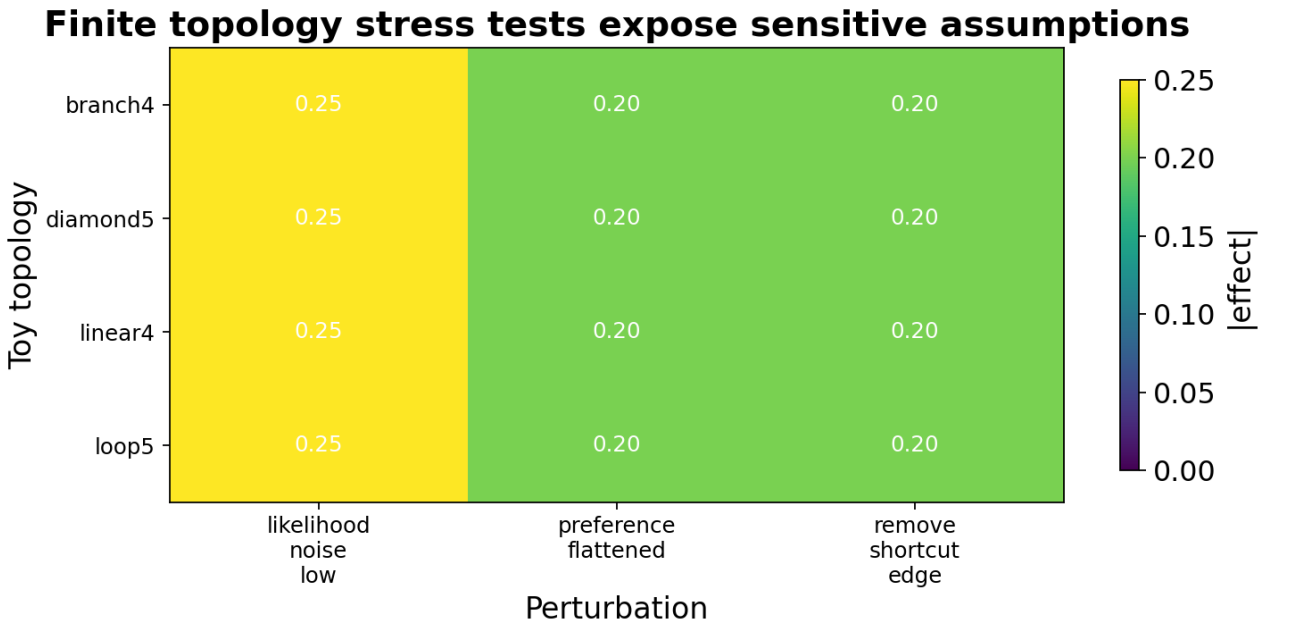


Figure 32: Causal-ablation heatmap over 36 source-backed rows joined to the sensitivity and uncertainty artifacts (all effects source-backed: true). This heatmap is an aggregated max-effect view of the 36 source rows: rows are toy graph topologies, columns are perturbation types, and each cell reports the maximum absolute deterministic effect of that intervention. The map shows which structural assumptions of the generative model the distillation outcome is sensitive to and which it is robust against inside this toy intervention matrix, flagging where the generated on-policy behavior would shift under declared model misspecification without asserting deployment-scale effects.

13.0.6 Appendix track: causal ablation

`causal_ablation` binds `output/data/causal_ablation_matrix.json` into the full sheaf appendix. Cells: 36. Complete grid: `true`.

The matrix is a finite teaching device: every row names a topology, a coupling value, a perturbation, a scalar effect, and the generated source row that made the effect admissible. It is not a claim about empirical interventions. It shows how an intervention-shaped table can be made falsifiable inside the sheaf: delete one perturbation cell or clear one deterministic flag and the grid gate fails before the manuscript can reuse the result.

`output/reports/ablation_sensitivity_report.json` then joins those ablation effects to the sensitivity and uncertainty artifacts. The report contributes 36 source-backed rows, with source-backed status `true`, so the appendix heatmap is a rendered view of validated JSON rather than a decorative restatement.

14 Supplementary material: reproducibility methodology

14.1 Compose contract

This standalone supplement documents the reproducibility methodology behind the rendered paper. The preceding full-coverage supplement (sec. 13) checks that the maximal appendix row can bind all registered fragment families; this section instead explains the operational contract that makes those fragments reproducible: where data are generated, how variables are hydrated, which validators run, and how failed gates block the PDF.

Each manifest row in `manuscript/sheaf/manifest.yaml` binds fragment tracks from `manuscript/sheaf/tracks.yaml`. A track supplies a renderer, compose order, label, and optional flag; the composer flattens the binding set into one Markdown section for PDF and web output. The machinery is generic, but the manuscript it assembles here argues a specific thesis: that on-policy distillation admits a finite-model active-inference reading when the variational objects are declared, so the composer must keep the analytical toy model, the pymdp rollout, the sequential-shift witness and sensitivity sweep, and the self-distillation literature mutually consistent about that scoped correspondence.

The operational claim is auditable binding: analytical, simulation, pymdp, visualization, Lean, GNN, ontology, scholarship, and optional media fragments attach to each IMRAD row under eq. 6 (**P** present, — unbound, **M** missing). This is an applied local-to-global consistency and composition-contract use of sheaf language in the spirit of cellular sheaves, sheaf-theoretic contracts, sheaf-signal-processing work, sensor-integration sheaves, semantic sheaving, applied compositionality, and reproducible computational research references [Curry, 2014, Speranzon et al., 2018, Robinson, 2014, 2017, Phillips, 2020, Fong and Spivak, 2019, Rosiak, 2022, Cox, 2026, Sandve et al., 2013, Wilkinson et al., 2016], but instantiated here as a finite manuscript artifact gate rather than as a public archive or release claim. Concretely, what this gate verifies is machine-executable provenance and version capture in the sense of standard reproducible-research practice [Sandve et al., 2013]; that discipline is necessary but not sufficient, since findable, accessible, interoperable, and reusable artifacts [Wilkinson et al., 2016] are not the same thing as end-to-end rerunnability or independent reproduction of the toy results by a third party, neither of which is claimed here. The same gate forces the teacher-student framing to remain coherent end to end: the Bernoulli-Ising free-energy analysis [Friston et al., 2006, 2009, Friston, 2010], the sophisticated-inference T-maze rollout [Parr et al., 2022, Da Costa et al., 2020], the sequential-shift witness and sensitivity sweep, and the on-policy distillation context [Agarwal et al., 2024, Lu and Thinking Machines Lab, 2025] each occupy their own track yet must agree on the variational posterior they describe.

14.2 Coverage and figures

fig. 33 summarizes 33 fragment types and their IMRAD bindings. Generated tables below list every track definition and section×track binding at compose time. The bindings span the full argument: the minimal-model demonstrations (analytical and pymdp tracks) and the scholarship track that situates them against the off-policy baseline [Hinton et al., 2015], the reverse-KL turn [Gu et al., 2024], and the 2026 self-distillation wave [Zhao et al., 2026, Shenfeld et al., 2026, Liu et al., 2026e].

The visualization layer is audited as data, not as decoration. `output/data/figure_source_map.json` binds every registered figure to source artifacts, source fields, validation gates, and explicit caption-claim contracts; `output/reports/figure_hash_manifest.json` records the declared rendered image bytes; and `output/reports/visualization_quality_audit.json` rechecks readability, nonblank pixels, source binding, caption-claim source fields, caption scope guardrails, cover wording, cover quantitative-free status, declared palette contrast, font-role floors, and absence of unregistered image artifacts. The negative controls mutate rows under green summaries and add stray image files, so a figure with an unreadable image, missing source, missing caption-claim field, unscoped empirical/production caption, inaccessible declared style token, stale cover equality claim, metric-dashboard cover language, or stale unregistered PNG cannot remain validated by a stale boolean.

The statistical layer follows the same rule. `output/data/firstprinciples/statistics_demo.json` is accepted only when its matched teacher/student entropy series, paired deltas, summaries, effect size, and seeded permutation metadata rederive from `output/data/firstprinciples/classroom.json` at validation time. This makes the classroom inferential paragraph a source-bound toy summary rather than a free-floating significance claim.

14.3 Compose commands

```
uv run python scripts/compose_manuscript.py
uv run python scripts/compose_manuscript.py --validate-only --strict
```

Each run emits `output/data/sheaf_coverage_matrix.json` and regenerates coverage artifacts. Partial compose (`--section`) is draft-only; the matrix always reflects the full manifest. Coverage totals appear on sec. 1; discussion scope is in sec. 11.

14.4 Law verification

`--validate-only --strict` runs the structural gate before any fragment is glued. Beyond per-cell coverage, it invokes the sheaf-law oracle (`verify_sheaf_laws`, `src/manuscript/sheaf/laws.py`), which checks 6 axioms — poset, presheaf functoriality, separation, gluing, typing, and compositionality — and reports 6/6 satisfied for the current manifest. A violation is raised as an error-level issue and aborts the build, so a malformed manifest (a section colliding on an output file, an off-chain block, a mistyped fragment, a fragment shared between sections) can never compose. The formal statements are in the formalism block below; the negative-control suite (`tests/test_sheaf_laws.py`) proves each check is falsifiable.

Stored summary flags are themselves never trusted at the final gate. Each generated artifact carries `all_*` aggregate booleans written by its producer; `validate_outputs` re-derives 62 of these aggregates from their own row data at read time (`src/gates/aggregate_rederivation.py`) and fails when a stored flag disagrees with its rows — including the vacuous case of a `true` flag over an empty row set. A mutated row under an untouched green summary therefore fails validation no matter what wrote it; the negative-control suite exercises exactly that lying case.

The semantic layer is separate from those structural laws. `output/data/sheaf_gluing_certificate.json` records cross-track symbols, typed claim evidence, artifact sources, and manuscript-variable restrictions; validation fails when the analytical, pymdp, GNN, ontology, Lean, visualization, or manuscript tracks disagree about a shared symbol or measured claim. This is where the correspondence is held honest at the symbol level: the coupling parameter and mutual information of the analytical toy, the cue-validity privileged-information channel of the T-maze, the two-agent classroom figures (privileged teacher belief entropy 0.247 nats versus the on-policy student's 0.347 nats, mean reverse-KL distillation signal 6.28 nats), and the sequential-shift witness (train loss 0.333 nats, induced test loss 0.409 nats, corrected test loss 0.096 nats, sensitivity loss reduction 0.313 nats) must all restrict consistently onto the shared variational-free-energy and reverse-KL symbols. The certificate keeps these numbers bound as a minimal-model demonstration of the teacher-student correspondence, not as claims about production LLMs. fig. 34 renders this gluing graph: the configured producers, the generated evidence artifacts, and the validation consumers that read each shared symbol.

14.4.1 Base poset and presheaf

The manuscript is modelled as a coverage sheaf over a finite base poset. Let the **base** P be the IMRAD blocks ordered as a chain,

$$\mathsf{Introduction} \prec \mathsf{Methods} \prec \mathsf{Results} \prec \mathsf{Discussion} \prec \mathsf{Appendix}, \tag{5}$$

with, in each block, a *group* node above its *section* nodes (written $G \sqsupseteq s$). P is therefore a finite poset (equivalently a finite Alexandrov space). Let $\mathcal{T}$ be the registered fragment-track set from `manuscript/sheaf/tracks.yaml`; each track $t \in \mathcal{T}$ carries a renderer $R(t)$, label $L(t)$, optional flag $O(t)$, and a strict compose-order index $\pi(t)$.

The **presheaf** $\mathcal{F}$ is a contravariant functor on P — $\mathcal{F} \colon P \to \mathbf{Set}$ with restriction maps along $\sqsupseteq$ — assigning to each composing section s its bound fragment set $\mathcal{F}(s) = \{\,(t, F_s(t)) : t \text{ bound in } s\,\}$, where $F_s : \mathcal{T} \rightharpoonup \mathbf{Path}$ is the section's partial binding map. Restriction along $G \sqsupseteq s$ is projection onto a section's own bindings; group nodes carry the empty assignment and do not compose.

The coverage cell is

$$B(s,t) \in \{\mathrm{P}, -, \mathrm{M}\} \tag{6}$$

derived from $F_s(t)$ and filesystem existence at compose time: **P** when a bound fragment exists, — when the track is unbound for that row, and **M** when a bound path is missing. The current regenerated matrix reports 95 present / 95 bound / 0 missing cells. Registry size: $|\mathcal{T}| = 33$ types across 17 IMRAD manifest rows (5 group rows, 12 composing sections).

14.4.2 Verified sheaf laws

What makes this presheaf a *sheaf* — rather than a bare incidence table — is that the composer's structural axioms are machine-checked. The oracle `verify_sheaf_laws` (`src/manuscript/sheaf/laws.py`) verifies 6 laws, and the regenerated build reports 6/6 satisfied:

1. **Poset.** The IMRAD blocks form the chain of eq. 5; compose order is monotone in block rank and every composing section's block carries a group row.
2. **Presheaf (functoriality).** Every bound track lies in $\mathcal{T}$; π is a strict total order; and each section's resolved track order is the monotone restriction of π (an explicit `track_order` override must be a permutation of the section's bound tracks).
3. **Separation (locality).** The map $s \mapsto$ output_name(s) is injective over composing sections: distinct locals glue to distinct global positions, so the global section is unique.
4. **Gluing.** Compose order is a linear extension of P — each block's rows are contiguous and strictly increasing in order — so the local fragments glue to a unique global manuscript in which every composing section appears exactly once.
5. **Typing.** Each binding $(t, F_s(t))$ is well-typed: $R(t)$ is a registered renderer and the fragment suffix lies in $R(t)$'s accepted suffix set. Generated renderers (`section_figures`, `layers_report`) synthesize their body and are explicitly type-exempt.
6. **Compositionality.** Every fragment file is private to one section (no path is bound twice), so global composition is the coproduct of the per-section bodies and is independent of inclusion order.

Each law is paired with a negative control in `tests/test_sheaf_laws.py` — a single mutation that breaks the law and is proven to be caught — so the gate binds the laws' *content*, not merely their shape. Under `--strict`, any violation is surfaced as an error-level manifest issue and aborts composition.

14.4.3 Scope (what is and is not claimed)

These laws verify the sheaf *axioms* on a finite base poset. They do **not** compute sheaf *cohomology* (H^0/H^1, Čech complexes, derived functors); "sheaf" here names the verified separation-and-gluing structure of a multi-track coverage assignment, not a cohomological invariant. The applied contracts reading is limited to the same finite local-to-global assembly discipline [Speranzon et al., 2018], not a claim that the manuscript instantiates a full systems-of-systems semantics. Formal track definitions and section×track bindings appear in the generated tables below.

Semantic gluing then checks agreement of the glued content: coverage counts, manuscript variables, typed claim predicates, pymdp mode/hash, Bernoulli GNN ontology, and SI T-maze GNN ontology. This certificate is a content-level audit over the same base, not an additional topological law.

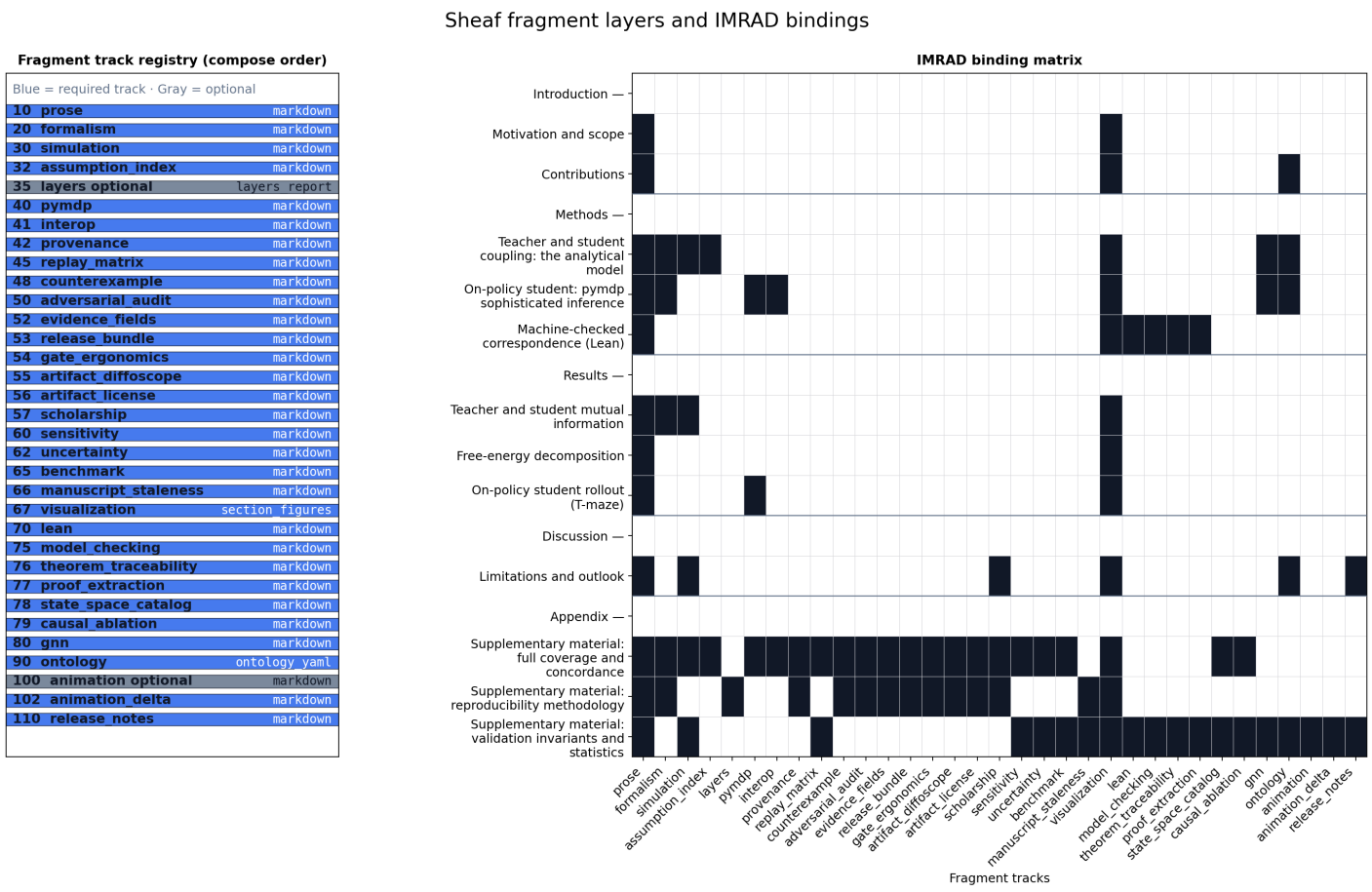


Figure 33: Sheaf layers overview. Left: the registry stack of 33 composable fragment types in compose order with their renderer ids, the ordered base over which the manuscript sheaf is assembled. Right: the IMRAD section-binding heatmap across 17 manifest rows (95 present / 95 bound / 0 missing). Together the panels show how heterogeneous local evidence – analytical, pymdp, and Lean fragments – is layered and bound section by section, the constructive mechanism by which the multi-track active-inference and on-policy-distillation argument is glued into a single coherent document.

The `provenance` fragment makes artifact lineage a live canonical sheaf track. The configured producer `generate_sheaf_tracks.py` writes `output/data/artifact_provenance.json`, which hashes 121 required toy artifacts and records producer scripts, source commit, deterministic seed fields, config digests, and 5 artifact bundles. Publication claims that depend on generated files must be traceable to this lineage table or to a narrower artifact-specific certificate.

The provenance claim is intentionally limited: every listed artifact exists, has a SHA-256 digest or an explicit cycle exclusion, is produced by a configured analysis script, and carries seed/config provenance (`121` seeded rows; all seeded flag `true`; bundle-complete flag `true`). A changed file, missing producer, or stale saved digest is a validation failure, not a prose warning.

The `counterexample` fragment records expected-failure fixtures as first-class evidence. `output/reports/counterexample_matrix.json` lists 23 negative controls that intentionally mutate ontology mappings, semantic certificates, graph-world trace agreement, typed claim evidence, replay rows, release parity, and provenance hashes.

The matrix is not an empirical result. It is a falsifiability ledger: each row names the gate that must fail and the test that proves the failure path remains live.

The `adversarial_audit` fragment makes expected failures part of the sheaf rather than an informal test note. `output/reports/adversarial_audit.json` records 23 known-bad rows and 0 known-bad rows passing; publication proceeds only when every row is documented as an expected failure and mapped to a gate.

The audit rows target the same failure modes as the semantic certificate: incomplete sweep cells, unnormalized uncertainty rows, interop field loss, stale certificate state, and empirical-scope leakage. The scope boundary remains toy-only: `toy_only_pass`.

The `evidence_fields` fragment indexes the exact artifact fields that support typed claims and hydrated manuscript tokens. `output/data/evidence_field_index.json` records 145 field rows, and the track passes only when every referenced JSONPath or dotted field is present (`true`).

The `release_bundle` fragment records whether the canonical deliverables exist before copying and whether copied root outputs

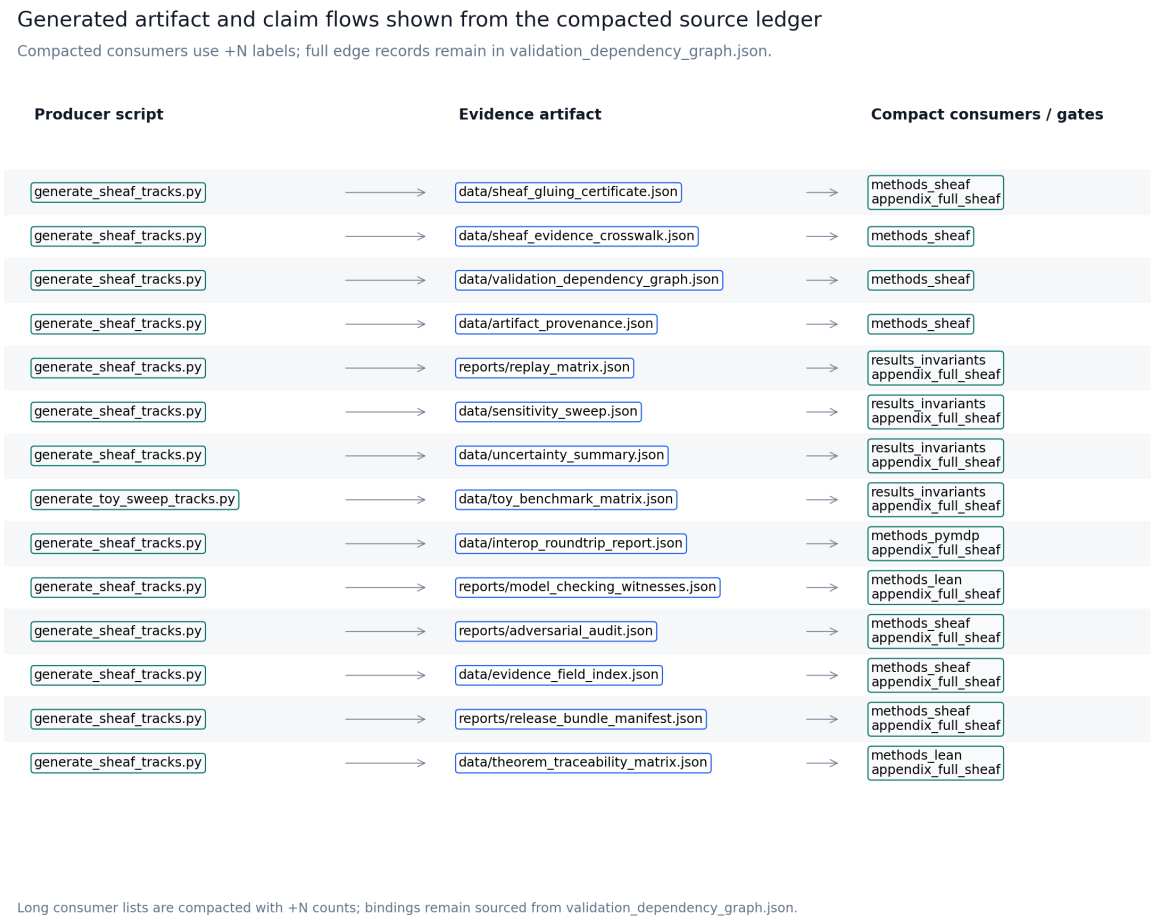


Figure 34: Semantic gluing graph tracing the dependency chain from configured analysis scripts (producers) through the generated evidence artifacts to the manuscript consumers and validation gates that close the multi-track sheaf certificate. Each edge records a declared provenance link, so the graph is the auditable trail showing that registered figure and variable dependencies are traced to declared producers and re-checked downstream. It is the operational embodiment of the sheaf gluing condition for this artifact contract: producers, artifacts, and consumers must agree along the registered edges before the assembled active-inference / on-policy-distillation argument is accepted. Long consumer lists are visually compacted with `+N` counts while remaining bound to `output/data/validation_dependency_graph.json`.

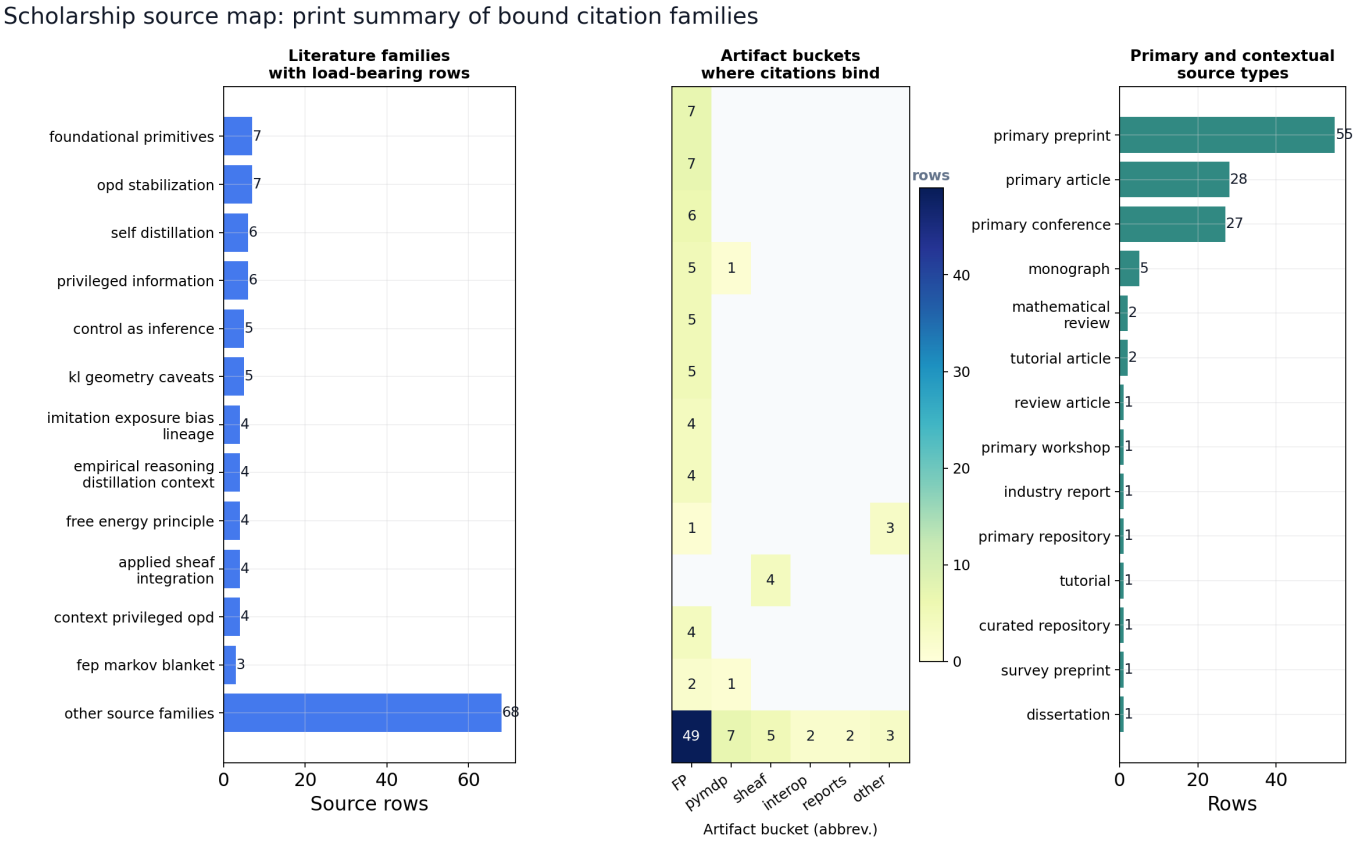


Figure 35: Condensed scholarship source map for 127 bibliography source rows across 127 method roles and 63 source families (connected status: true). The rendered figure is intentionally print-condensed: it shows the largest source families plus an aggregated long-tail row, where those families bind into generated artifact buckets, and the distribution of source kinds; the full row-level contract remains in `output/data/scholarship_source_matrix.json`. The map ties each external reference – on-policy distillation and active-inference literature alike – to a concrete place where the exemplar uses or tests it, evidencing load-bearing scholarship rather than decorative citation.

match or are explicitly deferred until the copy stage. `output/reports/release_bundle_manifest.json` tracks 39 required deliverables with source-present flag `true`.

The `gate_ergonomics` fragment turns validation commands into evidence rows. `output/data/validation_gate_index.json` records 27 gate rows, each naming required inputs and the negative-control surface that should fail closed.

Integration-audit sub-artifacts. Beyond the named sheaf tracks, `generate_integration_audit.py` emits a set of cross-cutting audit artifacts that are each enforced with a fail-closed negative control, so this section states what every one of them guarantees rather than leaving them as unexplained inventory rows. *Producer completeness* (`output/reports/producer_completeness.json`) requires every registered sheaf-track artifact to name a configured producer script; its `all_complete` flag is re-derived from the rows, so a registered artifact with a missing or unconfigured producer fails even if the stored flag was left true. *Token provenance* (`output/data/manuscript_token_provenance.json`) maps each hydrated double-brace token placeholder back to the artifact and field that produced it; the gate independently re-scans the manuscript and requires the rendered-token set, the provenance-key set, the per-row token set, and the live re-scan to coincide, so a deleted provenance row (a rendered token with no producer) or a phantom row (a provenance key that is never rendered) fails. *Claim-evidence audit* (`output/reports/claim_evidence_audit.json`) re-derives `all_claims_typed` per row, so any manuscript claim lacking a typed evidence binding or track set fails. *Scope-boundary audit* (`output/reports/scope_boundary_audit.json`) keeps every current claim inside the deterministic toy boundary and fails on any empirical or production scope leak. *Cross-track symbol table* (`output/data/cross_track_symbol_table.json`) is the table of shared symbols and the tracks that must agree on each; it backs the gluing certificate, which fails if two tracks bind different values to one symbol. *Evidence crosswalk* (`output/data/sheaf_evidence_crosswalk.json`) ties each typed claim to the evidence artifact and gate that back it; its schema gate fails closed on a malformed or inconsistent crosswalk, and its presence is enforced upstream by the producer-completeness check. *Validation dependency graph* (`output/data/validation_dependency_graph.json`) is the producer-to-artifact-to-consumer edge set that the semantic-gluing figure renders; manuscript validation fails on an unresolved dependency edge. *Reproducibility replay* (`output/data/reproducibility_replay.json`) records the end-to-end validation-spine replay distinct from the per-producer replay matrix, and its schema gate fails closed on a malformed replay record.

14.4.4 Artifact diffoscope track

The `artifact_diffoscope` track compares saved provenance hashes against live artifact hashes at the artifact root JSONPath. Its proof artifact is `output/reports/artifact_diffoscope.json`: it currently records 81 comparison rows, with equality status `true`.

14.4.5 Artifact license track

The `artifact_license` track classifies generated and project-source artifacts under the public project license boundary. Its audit artifact is `output/reports/artifact_license_audit.json`: it currently records 121 rows, with license-safe status `true`.

The `scholarship` fragment turns citations into an audited method surface rather than decorative bibliography. `output/data/scholarship_source_matrix.json` records 127 source rows across 127 method roles and 63 source families; fig. 35 renders the resulting source-to-artifact map. The row set connects foundational KL, variational-inference, model-compression, sequence-KD, and policy-distillation primitives [Kullback and Leibler, 1951, Jordan et al., 1999, Blei et al., 2017, Buciluǎ et al., 2006, Kim and Rush, 2016, Rusu et al., 2016, Czarnecki et al., 2019], foundational free-energy, predictive-coding, Markov-blanket, and active-inference references [Friston et al., 2006, 2009, Friston, 2010, 2013, Kirchhoff et al., 2018, Rao and Ballard, 1999, Buckley et al., 2017, Friston et al., 2017a,b, 2018, Millidge et al., 2021a, Da Costa et al., 2020, Friston et al., 2021a, Parr and Friston, 2019, Millidge et al., 2021b, Champion et al., 2024, Sajid et al., 2021a,b, de Vries et al., 2025, Parr et al., 2022, Smith et al., 2022, Tschantz et al., 2020a, Friston et al., 2021b, Aguilera et al., 2022, Parr et al., 2020], the sequential distribution-shift, behavioral-cloning, and distillation lineage [Pomerleau, 1989, Ross and Bagnell, 2010, Ross et al., 2011, Shimodaira, 2000, Sun et al., 2017, Bengio et al., 2015, Arora et al., 2022, Rohatgi et al., 2025, Pozzi et al., 2025, Hinton et al., 2015, Stanton et al., 2021, Gu et al., 2024, Agarwal et al., 2024, Yang et al., 2024, Ko et al., 2024, 2025, Wu et al., 2024, GX-Chen et al., 2025, Zelikman et al., 2022], reinforcement-learning/control-as-inference, MaxEnt-IRL, and preference-tilt bridges [Todorov, 2008, Toussaint, 2009, Ziebart et al., 2008, Levine, 2018, Abdolmaleki et al., 2018, Millidge et al., 2020a, O'Donoghue et al., 2020, Millidge et al., 2020b, Tschantz et al., 2020b, Haarnoja et al., 2018, Ouyang et al., 2022, Ziegler et al., 2019, Rafailov et al., 2023], privileged-information sources [Vapnik and Vashist, 2009, Lopez-Paz et al., 2016, Sharoni and Sabato, 2023, Cai et al., 2024, Penaloza et al., 2026a,b], recent self-distillation and entropy/hybrid OPD references [Zhao et al., 2026, Shenfeld et al., 2026, Hübotter et al., 2026, Liu et al., 2026e,d, Jin et al., 2026, Zhu et al., 2026b, Liu et al., 2026b, Oh et al., 2026, Xing et al., 2026, Jang et al., 2026, Ye et al., 2025], empirical reasoning-distillation and speculative-KD context [Qwen Team, 2025, Lu and Thinking Machines Lab, 2025, DeepSeek-AI, 2025, Xu et al., 2024], OPD landscape indexes [Liu, 2026, Song and Zheng, 2026, Zhu et al., 2026a, Ramos et al., 2026, Liu et al., 2026c], implementation, reproducibility, and notation anchors [Heins et al., 2022, Smékal and Friedman, 2023, Koudahl et al., 2023, Sandve et al., 2013, Wilkinson et al., 2016], applied sheaf sources [Curry, 2014, Speranzon et al., 2018, Robinson, 2014, 2017, Phillips, 2020, Fong and Spivak, 2019, Rosiak, 2022, Cox, 2026], and statistical-method reporting [Cohen, 1988] to the exact artifact or method role they support.

The validation claim is deliberately narrow: every row must have a bibliography entry with a DOI or URL, a manuscript citation, a registered sheaf track, a bound manifest section, an existing evidence artifact, and a claim-boundary statement. The hydrated flag `true` is therefore a source-traceability claim, not a claim that the toy results inherit empirical support from the cited literature.

The `manuscript_staleness` fragment closes the hydration loop. `output/reports/manuscript_staleness_report.json` checks 421 manuscript token bindings against the current generated variables after resolved markdown is written; the pass flag is `true`.

`output/reports/manuscript_hardcoded_variable_audit.json` then scans the source fragments for guarded generated values that appear as prose literals instead of double-brace manuscript-variable placeholders. It guards 115 formatted token values and records 0

hard-coded-value issues; the pass flag is `true`.

This is a publication-systems claim, not a domain result. A stale hydrated value, unresolved token, hard-coded generated value, or missing resolved section becomes a validation failure before PDF or web outputs are accepted.

14.5 Sheaf fragment track registry

Compose order and renderer bindings from `manuscript/sheaf/tracks.yaml`.

Order	Track id	Label	Renderer	Optional
10	`prose`	Narrative prose	`markdown`	No
20	`formalism`	Mathematical formalism	`markdown`	No
30	`simulation`	Analytical simulation notes	`markdown`	No
32	`assumption_index`	Analytical assumption index	`markdown`	No
35	`layers`	Sheaf layers tables	`layers_report`	Yes
40	`pymdp`	pymdp harness artifacts	`markdown`	No
41	`interop`	GNN/ontology/JSON interop checks	`markdown`	No
42	`provenance`	Artifact provenance and bundle lineage spine	`markdown`	No
45	`replay_matrix`	Deterministic replay matrix	`markdown`	No
48	`counterexample`	Expected-failure counterexamples	`markdown`	No
50	`adversarial_audit`	Adversarial audit matrix	`markdown`	No
52	`evidence_fields`	Evidence field index	`markdown`	No
53	`release_bundle`	Release bundle parity manifest	`markdown`	No
54	`gate_ergonomics`	Validation gate ergonomics	`markdown`	No
55	`artifact_diffoscope`	Artifact diffoscope	`markdown`	No
56	`artifact_license`	Artifact license audit	`markdown`	No
57	`scholarship`	Source-backed scholarship matrix	`markdown`	No
60	`sensitivity`	Toy sensitivity sweep	`markdown`	No
62	`uncertainty`	Toy uncertainty summaries	`markdown`	No
65	`benchmark`	Compact toy benchmark matrix	`markdown`	No
66	`manuscript_staleness`	Hydrated manuscript staleness report	`markdown`	No
67	`visualization`	Figure references	`section_figures`	No
70	`lean`	Lean boundary fragment	`markdown`	No
75	`model_checking`	Finite-state model checking witnesses	`markdown`	No
76	`theorem_traceability`	Lean theorem traceability matrix	`markdown`	No
77	`proof_extraction`	Lean proof extraction index	`markdown`	No
78	`state_space_catalog`	Finite state-space catalog	`markdown`	No
79	`causal_ablation`	Deterministic causal ablation matrix	`markdown`	No
80	`gnn`	GNN notation fragment	`markdown`	No

Order	Track id	Label	Renderer	Optional
90	`ontology`	Active Inference Ontology bindings	`ontology_yaml`	No
100	`animation`	Animation fragment	`markdown`	Yes
102	`animation_delta`	Animation frame-delta manifest	`markdown`	No
110	`release_notes`	Release notes evidence	`markdown`	No

Track count: 33 registered fragment types.

14.6 IMRAD binding matrix

Section rows versus fragment track columns. **P** = present (bound and file exists); — = absent (not bound); **M** = missing (bound, file absent).

Section	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Introduction (group)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Motivation and scope	P	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	P	—	—	—	—	—	—	—	—	—	—	—
Contributions	P	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	P	—	—	—	—	—	—	—	P	—	—	—
Methods (group)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Teacher and student coupling: the analytical model	P	P	P	P	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	P	—	—	—	—	—	—	P	P	—	—	—
On-policy student: pymdp sophisticated inference	P	P	—	—	—	P	P	—	—	—	—	—	—	—	—	—	—	—	—	—	—	P	—	—	—	—	—	—	P	P	—	—	—
Machine-checked correspondence (Lean)	P	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	P	P	P	P	P	—	—	—	—	—	—	—
Results (group)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

Section	proof	formalism	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Teacher and student mutual information	P	P	P	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	P	—	—	—	—	—	—	—	—	—	—	—
Free energy decomposition	P	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	P	—	—	—	—	—	—	—	—	—	—	—
On-policy student rollout (T-maze)	P	—	—	—	—	P	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	P	—	—	—	—	—	—	—	—	—	—	—
Discussion (group)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Limitations and outlook	P	—	P	—	—	—	—	—	—	—	—	—	—	—	—	—	P	—	—	—	—	P	—	—	—	—	—	—	—	P	—	—	P
Appendix (group)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Supplementary material: full coverage and concordance	P	P	P	P	—	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	—	P	—	—	—	—	P	P	—	—	—	—	—
Supplementary material: reproducibility methodology	P	P	—	—	P	—	—	P	—	P	P	P	P	P	P	P	P	—	—	—	P	P	—	—	—	—	—	—	—	—	—	—	—

Section																																	
Supplementary material: validation invariants and statistics	P	—	P	—	—	—	—	—	P	—	—	—	—	—	—	—	—	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P

Totals: 95 present / 95 bound / 0 missing.

Symbol	Coverage color	Meaning
P	Black	Track **present** (bound and fragment exists)
—	White	**Absent** (not bound for this section)
M	Gray	**Missing** (bound but fragment file absent)

14.7 Section-track status

Generated status for the current manuscript sheaf, summarized per composable section.

Section	IMRAD	Bound	Present	Missing	Status
Motivation and scope	introduction	2	2	0	`fully_sheafed`
Contributions	introduction	3	3	0	`fully_sheafed`
Teacher and student coupling: the analytical model	methods	7	7	0	`fully_sheafed`
On-policy student: pymdp sophisticated inference	methods	7	7	0	`fully_sheafed`
Machine-checked correspondence (Lean)	methods	6	6	0	`fully_sheafed`
Teacher and student mutual information	results	4	4	0	`fully_sheafed`
Free-energy decomposition	results	2	2	0	`fully_sheafed`
On-policy student rollout (T-maze)	results	3	3	0	`fully_sheafed`
Limitations and outlook	discussion	6	6	0	`fully_sheafed`
Supplementary material: full coverage and concordance	appendix	22	22	0	`fully_sheafed`
Supplementary material: reproducibility methodology	appendix	14	14	0	`fully_sheafed`

Section	IMRAD	Bound	Present	Missing	Status
Supplementary material: validation invariants and statistics	appendix	19	19	0	fully_sheafed

Section status: 12 / 12 composable sections fully sheafed; 0 required bound fragments missing.

14.8 Track status

Track	Renderer	Bound sections	Present	Missing	Claims	Status
prose	markdown	12	12	0	0	complete
formalism	markdown	5	5	0	0	complete
simulation	markdown	5	5	0	29	complete
assumption_index	markdown	2	2	0	1	complete
layers	layers_report	1	1	0	1	complete
pymdp	markdown	3	3	0	24	complete
interop	markdown	2	2	0	6	complete
provenance	markdown	2	2	0	12	complete
replay_matrix	markdown	2	2	0	3	complete
counterexample	markdown	2	2	0	2	complete
adversarial_audit	markdown	2	2	0	11	complete
evidence_fields	markdown	2	2	0	2	complete
release_bundle	markdown	2	2	0	5	complete
gate_ergonomics	markdown	2	2	0	5	complete
artifact_diffoscope	markdown	2	2	0	1	complete
artifact_license	markdown	2	2	0	1	complete
scholarship	markdown	3	3	0	12	complete
sensitivity	markdown	2	2	0	10	complete
uncertainty	markdown	2	2	0	6	complete
benchmark	markdown	2	2	0	3	complete
manuscript_staleness	markdown	2	2	0	1	complete
visualization	section_figures	12	12	0	16	complete
lean	markdown	2	2	0	8	complete
model_checking	markdown	2	2	0	7	complete
theorem_traceability	markdown	2	2	0	3	complete
proof_extraction	markdown	2	2	0	2	complete
state_space_catalog	markdown	2	2	0	2	complete
causal_ablation	markdown	2	2	0	2	complete
gnn	markdown	3	3	0	5	complete
ontology	ontology_yaml	5	5	0	7	complete
animation	markdown	1	1	0	2	complete
animation_delta	markdown	1	1	0	1	complete
release_notes	markdown	2	2	0	2	complete

Status cells: 561 section-track cells.

14.9 Render and logging summary

Event	Component	Output	Status	Detail
registry_loaded	sheaf.registry	registered_tracks	ok	33 tracks
manifest_loaded	sheaf.manifest	manifest_sections	ok	17 sections
coverage_matrix_built	sheaf.coverage	output/data/sheaf_coverage_matrix.json	ok	95 present cells
section_status_matrix_built	sheaf.status	output/data/sheaf_section_status_matrix.json	ok	561 section-track cells
layers_renderer_bound	sheaf.layers_report	manuscript/19_supplement_reproducibility.md	ok	methods sheaf layer tables
semantic_artifacts_indexed	sheaf.semantic	output/data/validation_dependency_graph.json	ok	121 artifact producer rows
validation_gates_indexed	gates	output/data/validation_gate_index.json	ok	3 gate groups
manuscript_sections_composed	sheaf.compose	manuscript/*.md	ok	16 composed markdown files

Render events: 8.

14.10 Evidence crosswalk

Claim	Artifact	Producer	Gates
sheaf_registry	manuscript/sheaf/tracks.yaml	manual	validate_outputs
sheaf_manifest	manuscript/sheaf/manifest.yaml	manual	validate_outputs
sheaf_coverage_config	manuscript/sheaf/coverage.yaml	manual	validate_outputs
sheaf_coverage_matrix	output/data/sheaf_coverage_matrix.json	generate_figures.py	validate_outputs, validate_manuscript
sheaf_gluing_certificate	output/data/sheaf_gluing_certificate.json	generate_sheaf_tracks.py	validate_manuscript, validate_outputs
sheaf_evidence_crosswalk	output/data/sheaf_evidence_crosswalk.json	generate_sheaf_tracks.py	validate_manuscript, validate_outputs
evidence_field_index	output/data/evidence_field_index.json	generate_sheaf_tracks.py	validate_outputs, validate_manuscript
validation_dependency_graph	output/data/validation_dependency_graph.json	generate_sheaf_tracks.py	validate_manuscript, validate_outputs

Claim rows: 145 typed evidence claims.

14.11 Artifact producer graph

Artifact	Producer	Configured	Consumers
output/data/analysis_statistics.json	compute_statistics.py	Yes	results_si_tmaze, results_invariants
output/data/analytical_assumption_index.json	generate_toy_sweep_tracks.py	Yes	methods_analytical, appendix_full_sheaf
output/data/analytical_observable_sweep.json	generate_toy_sweep_tracks.py	Yes	results_invariants, appendix_full_sheaf
output/data/animation_frame_deltas.json	render_animation.py	Yes	appendix_full_sheaf
output/data/artifact_provenance.json	generate_sheaf_tracks.py	Yes	methods_sheaf
output/data/causal_ablation_matrix.json	generate_toy_sweep_tracks.py	Yes	results_invariants, appendix_full_sheaf

Artifact	Producer	Configured	Consumers
output/data/cross_track_symbol_table.json	generate_integration_audit.py	Yes	methods_sheaf, appendix_full_sheaf
output/data/evidence_field_index.json	generate_sheaf_tracks.py	Yes	methods_sheaf, appendix_full_sheaf
output/data/figure_source_map.json	generate_integration_audit.py	Yes	methods_sheaf, appendix_full_sheaf
output/data/firstprinciples/active_selection_demo.json	generate_firstprinciples.py	Yes	results_free_energy
output/data/firstprinciples/active_selection_general_demo.json	generate_firstprinciples.py	Yes	results_free_energy
output/data/firstprinciples/benchmark_table.md	generate_firstprinciples.py	Yes	appendix_full_sheaf
output/data/firstprinciples/classroom.json	generate_firstprinciples.py	Yes	intro_motivation, results_si_tmaze, discussion_outlook
output/data/firstprinciples/correspondence_map.json	generate_firstprinciples.py	Yes	intro_contributions, methods_analytical, methods_sheaf, discussion_outlook
output/data/firstprinciples/correspondence_table.md	generate_firstprinciples.py	Yes	methods_sheaf, appendix_full_sheaf
output/data/firstprinciples/divergence_demo.json	generate_firstprinciples.py	Yes	methods_analytical, discussion_outlook
output/data/firstprinciples/empirical_benchmark.json	generate_firstprinciples.py	Yes	discussion_outlook, appendix_full_sheaf
output/data/firstprinciples/exposure_bias_demo.json	generate_firstprinciples.py	Yes	intro_motivation, methods_pymdp, discussion_outlook
output/data/firstprinciples/opd_taxonomy.json	generate_firstprinciples.py	Yes	intro_motivation, methods_sheaf, discussion_outlook
output/data/firstprinciples/precision_ledger_demo.json	generate_firstprinciples.py	Yes	results_free_energy
output/data/firstprinciples/privilege_sweep.json	generate_firstprinciples.py	Yes	results_si_tmaze, appendix_full_sheaf
output/data/firstprinciples/reward_tilting_demo.json	generate_firstprinciples.py	Yes	methods_analytical, discussion_outlook
output/data/firstprinciples/sdpg_demo.json	generate_firstprinciples.py	Yes	methods_analytical, discussion_outlook
output/data/firstprinciples/sequential_selection_demo.json	generate_firstprinciples.py	Yes	results_free_energy
output/data/firstprinciples/sequential_shift.json	generate_firstprinciples.py	Yes	results_si_tmaze, discussion_outlook
output/data/firstprinciples/sequential_shift_sensitivity.json	generate_firstprinciples.py	Yes	results_si_tmaze, discussion_outlook
output/data/firstprinciples/si_bridge_demo.json	generate_firstprinciples.py	Yes	results_free_energy
output/data/firstprinciples/statistics_demo.json	generate_firstprinciples.py	Yes	results_invariants, appendix_full_sheaf
output/data/firstprinciples/taxonomy_table.md	generate_firstprinciples.py	Yes	methods_sheaf, appendix_full_sheaf
output/data/gnn_roundtrip_report.json	generate_formal_interop_tracks.py	Yes	methods_pymdp, appendix_full_sheaf
output/data/interop_roundtrip_report.json	generate_sheaf_tracks.py	Yes	methods_pymdp, appendix_full_sheaf

Artifact	Producer	Configured	Consumers
output/data/manuscript_evidence_tables.json	generate_integration_audit.py	Yes	methods_sheaf, appendix_full_sheaf
output/data/manuscript_token_provenance.json	generate_integration_audit.py	Yes	methods_sheaf, appendix_full_sheaf
output/data/manuscript_variables.json	z_generate_manuscript_variables.py	Yes	methods_sheaf, appendix_full_sheaf
output/data/ontology_alias_index.json	generate_formal_interop_tracks.py	Yes	methods_pymdp, appendix_full_sheaf
output/data/ontology_profile_matrix.json	generate_formal_interop_tracks.py	Yes	methods_pymdp, appendix_full_sheaf
output/data/parameter_sweep.csv	run_analytical_sweep.py	Yes	methods_analytical, results_mi_sweep
output/data/proof_dependency_graph.json	generate_sheaf_tracks.py	Yes	methods_lean, appendix_full_sheaf
output/data/proof_extraction_index.json	generate_formal_interop_tracks.py	Yes	methods_lean, appendix_full_sheaf
output/data/pymdp_policy_posterior_grid.json	simulate_si_tmaze.py	Yes	methods_pymdp, appendix_full_sheaf
output/data/scholarship_source_matrix.json	generate_sheaf_tracks.py	Yes	methods_sheaf, appendix_full_sheaf
output/data/sensitivity_sweep.json	generate_sheaf_tracks.py	Yes	results_invariants, appendix_full_sheaf
output/data/sheaf_coverage_matrix.json	generate_figures.py	Yes	methods_sheaf, appendix_full_sheaf
output/data/sheaf_evidence_crosswalk.json	generate_sheaf_tracks.py	Yes	methods_sheaf
output/data/sheaf_gluing_certificate.json	generate_sheaf_tracks.py	Yes	methods_sheaf, appendix_full_sheaf
output/data/sheaf_section_status_matrix.json	generate_sheaf_tracks.py	Yes	methods_sheaf, appendix_full_sheaf
output/data/si_efe_terms.json	generate_toy_sweep_tracks.py	Yes	results_invariants, appendix_full_sheaf
output/data/si_graph_world_summary.json	simulate_si_graph_world.py	Yes	methods_pymdp, results_si_tmaze
output/data/si_graph_world_topology_sweep.json	generate_toy_sweep_tracks.py	Yes	results_invariants, appendix_full_sheaf
output/data/si_graph_world_topology_traces.json	generate_toy_sweep_tracks.py	Yes	results_invariants, appendix_full_sheaf
output/data/si_graph_world_trace.json	simulate_si_graph_world.py	Yes	methods_pymdp, results_si_tmaze, appendix_full_sheaf
output/data/si_policy_comparison.json	simulate_si_tmaze.py	Yes	methods_pymdp, results_si_tmaze
output/data/si_policy_grid.json	generate_toy_sweep_tracks.py	Yes	results_invariants, appendix_full_sheaf
output/data/si_tmaze_model_matrices.json	simulate_si_tmaze.py	Yes	methods_pymdp, results_si_tmaze, appendix_full_sheaf
output/data/si_tmaze_summary.json	simulate_si_tmaze.py	Yes	methods_pymdp, results_si_tmaze
output/data/si_tmaze_trace.json	simulate_si_tmaze.py	Yes	methods_pymdp, results_si_tmaze
output/data/state_space_catalog.json	generate_toy_sweep_tracks.py	Yes	results_invariants, appendix_full_sheaf
output/data/state_transition_table.json	generate_sheaf_tracks.py	Yes	results_invariants, appendix_full_sheaf
output/data/theorem_traceability_matrix.json	generate_sheaf_tracks.py	Yes	methods_lean, appendix_full_sheaf
output/data/toy_benchmark_matrix.json	generate_toy_sweep_tracks.py	Yes	results_invariants, appendix_full_sheaf

Artifact	Producer	Configured	Consumers
output/data/track_improvement_scope.json	generate_sheaf_tracks.py	Yes	methods_sheaf, appendix_full_sheaf
output/data/uncertainty_summary.json	generate_sheaf_tracks.py	Yes	results_invariants, appendix_full_sheaf
output/data/validation_dependency_graph.json	generate_sheaf_tracks.py	Yes	methods_sheaf
output/data/validation_gate_index.json	generate_integration_audit.py	Yes	methods_sheaf, appendix_full_sheaf
../figures/si_belief_trajectory.gif	render_animation.py	Yes	appendix_full_sheaf
output/reports/ablation_sensitivity_report.json	generate_sheaf_tracks.py	Yes	results_invariants, appendix_full_sheaf
output/reports/adversarial_audit.json	generate_sheaf_tracks.py	Yes	methods_sheaf, appendix_full_sheaf
output/reports/artifact_diffoscope.json	generate_integration_audit.py	Yes	methods_sheaf, appendix_full_sheaf
output/reports/artifact_license_audit.json	generate_integration_audit.py	Yes	methods_sheaf, appendix_full_sheaf
output/reports/blocked_scope_manifest.json	generate_sheaf_tracks.py	Yes	methods_sheaf, discussion_outlook, appendix_full_sheaf
output/reports/claim_evidence_audit.json	generate_integration_audit.py	Yes	methods_sheaf, appendix_full_sheaf
output/reports/counterexample_matrix.json	generate_sheaf_tracks.py	Yes	methods_sheaf
output/reports/figure_hash_manifest.json	generate_integration_audit.py	Yes	methods_sheaf, appendix_full_sheaf
output/reports/gnn_lint_report.json	generate_formal_interop_tracks.py	Yes	methods_pymdp, appendix_full_sheaf
output/reports/graph_world_invariants.json	generate_toy_sweep_tracks.py	Yes	results_invariants, appendix_full_sheaf
output/reports/invariants.json	run_analytical_sweep.py	Yes	results_invariants
output/reports/lean_graph_world_inventory.json	generate_formal_interop_tracks.py	Yes	methods_lean, appendix_full_sheaf
output/reports/lean_theorem_inventory.json	generate_formal_interop_tracks.py	Yes	methods_lean, appendix_full_sheaf
output/reports/manuscript_hardcoded_variable_audit.json	generate_integration_audit.py	Yes	methods_sheaf, appendix_full_sheaf
output/reports/manuscript_staleness_report.json	z_generate_manuscript_variables.py	Yes	methods_sheaf, appendix_full_sheaf
output/reports/model_checking_witnesses.json	generate_sheaf_tracks.py	Yes	methods_lean, appendix_full_sheaf
output/reports/producer_completeness.json	generate_integration_audit.py	Yes	methods_sheaf, appendix_full_sheaf
output/reports/pymdp_runtime_diagnostics.json	simulate_si_tmaze.py	Yes	methods_pymdp, appendix_full_sheaf
output/reports/release_attestation.json	generate_sheaf_tracks.py	Yes	discussion_outlook, appendix_full_sheaf
output/reports/release_bundle_manifest.json	generate_sheaf_tracks.py	Yes	methods_sheaf, appendix_full_sheaf
output/reports/release_notes_evidence.json	generate_integration_audit.py	Yes	discussion_outlook, appendix_full_sheaf
output/reports/replay_matrix.json	generate_sheaf_tracks.py	Yes	results_invariants, appendix_full_sheaf
output/reports/reproducibility_replay.json	generate_validation_spine.py	Yes	results_invariants
output/reports/scope_boundary_audit.json	generate_integration_audit.py	Yes	methods_sheaf, appendix_full_sheaf

Artifact	Producer	Configured	Consumers
output/reports/sheaf_render_log.json	generate_sheaf_tracks.py	Yes	methods_sheaf, appendix_full_sheaf
output/reports/si_invariants.json	simulate_si_tmaze.py	Yes	results_si_tmaze
output/reports/si_tmaze_run_report.json	simulate_si_tmaze.py	Yes	results_si_tmaze
output/reports/stale_artifact_report.json	generate_integration_audit.py	Yes	methods_sheaf, appendix_full_sheaf
output/reports/visualization_quality_audit.json	generate_integration_audit.py	Yes	methods_sheaf, appendix_full_sheaf

Producer issues: 0.

14.12 Semantic gluing restrictions

Restriction	Value
Coverage missing	0
Policy comparison rows	2
Policy grid complete	True
Policy posterior rows	14
Policy posterior normalized	True
Runtime unexpected warnings	0
Graph-world trace agrees	True
Animation frames	4
Lean all proved	True
GNN ontology ok	True
Configured producers ok	True
Semantic certificate ok	not evaluated
Dependency edges ok	True
Track scope complete	True
Empirical adapter blocked	True
Provenance bundles complete	True
Replay rows matched	True
Sensitivity complete	True
Uncertainty normalized	True
Evidence fields mapped	True
Release bundle sources present	True
Theorem traceability linked	True
Gate ergonomics indexed	True
Interop lossless	True
Scope toy-only	True

14.13 Track improvement scope

Track	Status	Current proof	Next artifact	Gate	Negative control
adversarial_audit	live	output/reports/adversarial_audit.json	output/reports/adversarial_audit.json	validate_outputs, validate_manuscript	adversarial_known_bad_pas
animation	optional	../figures/si_belief_trajectory.gif	../figures/si_belief_trajectory.gif	validate_outputs	missing_fragment_coverage
animation_delta	live	output/data/animation_frame_deltas.json	output/data/animation_frame_deltas.json	validate_outputs, validate_manuscript	missing_fragment_coverage
artifact_diffoscope	live	output/reports/artifact_diffoscope.json	output/reports/artifact_diffoscope.json	validate_outputs, validate_manuscript	artifact_diffoscope_missed_

Track	Status	Current proof	Next artifact	Gate	Negative control
artifact_license	live	output/reports/artifact_license_audit.json	output/reports/artifact_license_audit.json	validate_outputs, validate_manuscript	artifact_license_unsafe_arti
assumption_index	live	output/data/analytical_assumption_index.json	output/data/analytical_assumption_index.json	validate_outputs, validate_manuscript	missing_fragment_coverage
benchmark	live	output/data/toy_benchmark_matrix.json	output/data/toy_benchmark_matrix.json	validate_outputs	missing_fragment_coverage
causal_ablation	live	output/data/causal_ablation_matrix.json	output/data/causal_ablation_matrix.json	validate_outputs, validate_manuscript	causal_ablation_missing_ce
counterexample	live	output/reports/counterexample_matrix.json	output/reports/counterexample_matrix.json	validate_outputs, validate_manuscript	known_bad_counterexample
evidence_fields	live	output/data/evidence_field_index.json	output/data/evidence_field_index.json	validate_outputs, validate_manuscript	missing_typed_claim
formalism	live	manuscript/sheaf/manifest.yaml	manuscript/sheaf/manifest.yaml	validate_manuscript	missing_fragment_coverage
gate_ergonomics	live	output/data/validation_gate_index.json	output/data/validation_gate_index.json	validate_outputs, validate_manuscript	gate_ergonomics_unindexed
gnn	live	output/reports/gnn_lint_report.json	output/reports/gnn_lint_report.json	validate_outputs	missing_fragment_coverage
interop	live	output/data/interop_roundtrip_report.json	output/data/interop_roundtrip_report.json	validate_outputs	interop_shape_loss
layers	optional	output/data/sheaf_coverage_matrix.json	output/data/sheaf_coverage_matrix.json	validate_outputs, validate_manuscript	missing_fragment_coverage
lean	live	output/reports/lean_theorem_inventory.json	output/reports/lean_theorem_inventory.json	validate_outputs	missing_fragment_coverage
manuscript_staleness	live	output/reports/manuscript_staleness_report.json	output/reports/manuscript_staleness_report.json	validate_outputs, validate_manuscript	missing_fragment_coverage
model_checking	live	output/reports/model_checking_witnesses.json	output/reports/model_checking_witnesses.json	validate_outputs	missed_model_checking_co
ontology	live	output/data/ontology_profile_matrix.json	output/data/ontology_profile_matrix.json	validate_outputs	missing_fragment_coverage
proof_extraction	live	output/data/proof_extraction_index.json	output/data/proof_extraction_index.json	validate_outputs, validate_manuscript	proof_extraction_missing_s
prose	live	manuscript/sheaf/manifest.yaml	manuscript/sheaf/manifest.yaml	validate_manuscript	missing_fragment_coverage
provenance	live	output/data/artifact_provenance.json	output/data/artifact_provenance.json	validate_manuscript, validate_outputs	missing_sheaf_track_produ
pymdp	live	output/data/si_policy_comparison.json	output/data/si_policy_comparison.json	validate_outputs	missing_fragment_coverage
release_bundle	live	output/reports/release_bundle_manifest.json	output/reports/release_bundle_manifest.json	validate_outputs, validate_manuscript	release_bundle_parity_failu
release_notes	live	output/reports/release_notes_evidence.json	output/reports/release_notes_evidence.json	validate_outputs, validate_manuscript	release_notes_claim_failed_

Track	Status	Current proof	Next artifact	Gate	Negative control
replay_matrix	live	output/reports/replay_matrix.json	output/reports/replay_matrix.json	validate_outputs, validate_manuscript	replay_mismatch
scholarship	live	output/data/scholarship_source_matrix.json	output/data/scholarship_source_matrix.json	validate_outputs, validate_manuscript	missing_scholarship_source_
sensitivity	live	output/data/sensitivity_sweep.json	output/data/sensitivity_sweep.json	validate_outputs	missing_sensitivity_cell
simulation	live	output/data/analytical_observable_sweep.json	output/data/analytical_observable_sweep.json	validate_outputs	missing_fragment_coverage
state_space_catalog	live	output/data/state_space_catalog.json	output/data/state_space_catalog.json	validate_outputs, validate_manuscript	state_space_catalog_missin
theorem_traceability	live	output/data/theorem_traceability_matrix.json	output/data/theorem_traceability_matrix.json	validate_outputs, validate_manuscript	theorem_traceability_unlink
uncertainty	live	output/data/uncertainty_summary.json	output/data/uncertainty_summary.json	validate_outputs	unnormalized_uncertainty_
visualization	live	output/data/figure_source_map.json	output/data/figure_source_map.json	validate_outputs, validate_manuscript	missing_fragment_coverage
empirical_adapter	blocked	output/reports/blocked_scope_manifest.json	output/data/empirical_adapter_manifest.json	blocked_scope_manifest.all_blocked	empirical claim appears without manifest
private_or_restricted_data	blocked	output/reports/blocked_scope_manifest.json	output/data/private_data_provenance_manifest.json	blocked_scope_manifest.all_blocked	private data artifact appears without provenance manifest
network_dependent_research	blocked	output/reports/blocked_scope_manifest.json	output/data/network_replay_manifest.json	blocked_scope_manifest.all_blocked	network-derived claim appears without replay manifest
llm_generated_evidence	blocked	output/reports/blocked_scope_manifest.json	output/data/llm_evidence_audit.json	blocked_scope_manifest.all_blocked	LLM-generated evidence appears as a validation source
non_toy_model_claims	blocked	output/reports/blocked_scope_manifest.json	output/data/non_toy_model_scope_manifest.json	blocked_scope_manifest.all_blocked	non-toy result claim appears outside future-only scope

Improvement rows: 38.

15 Supplementary material: validation invariants and statistics

Because the central claim is a formal correspondence rather than a metaphor, every quantity that instantiates the teacher–student mapping is guarded by an invariant that runs before PDF rendering (sec. 5). These gates assert that the analytical free energy, the coupling-mediated mutual information $I(\lambda)$, and the reverse- versus forward-KL limits behave as the distillation objective predicts, and that the on-policy student rollout reported by the pymdp harness reproduces the privileged-information advantage. On a clean checkout **16 / 16** checks pass in the merged validation report, which records the sophisticated-inference simulation invariants whenever the pymdp harness ran (sec. 10).

The full registry, binding-matrix, and track-status layer tables appear once, in the reproducibility supplement (sec. 14); this section reports only the validation statistics layered on top of them. fig. 36 lists each analytical and simulation gate; a failure on any correspondence check blocks publication artifacts, so a broken claim about the variational-posterior-as-student mapping cannot pass silently. This validation supplement follows the reproducibility-methodology supplement (sec. 14) because the invariant counts depend on the same hydration and source-map boundary: the numbers are generated first, injected second, and only then rendered. As with the rest of this work, the guarantees are scoped to the analytical Bernoulli–Ising model and the T-maze rollout: they check that these minimal artifacts instantiate the on-policy-distillation correspondence faithfully, not that the same numbers hold for production-scale language-model distillation.

Beyond the binary pass/fail gates, the privileged-information asymmetry that defines the teacher–student Markov blanket admits a quantitative inferential test. Under the correspondence, the teacher conditions its generative model on privileged context the student cannot observe, so the teacher should hold a sharper posterior — lower belief entropy — than the on-policy student that must generate its own observations and update from them. We measure this gap directly in the two-agent pymdp classroom: `statistics_demo.json` is derived from the per-decision belief-entropy series persisted in `classroom.json`, giving 4 matched per-decision teacher/student pairs, and we report interval, effect size, and permutation p rather than a single point estimate. The validator rederives the entropy series, paired deltas, summaries, and test metadata from that classroom artifact before the manuscript can render, so the paragraph is bound to the measured toy rows rather than to a copied summary flag. By construction the mean paired difference equals the classroom's reported entropy gap. The privileged-teacher advantage is 0.100 nats (student minus teacher belief entropy), with a bootstrap confidence interval of [-0.098, 0.422] nats. At this sample size the raw data say more than any interval, so we print all 4 paired student-minus-teacher deltas directly: +0.000, +0.595, -0.098, -0.098 nats. The standardized effect size is Cohen's $d = 0.28$ using Cohen's d pooled standardized mean difference, student minus teacher [Cohen, 1988], and two-sided paired permutation test on mean student-minus-teacher entropy with 5000 seeded permutations returns $p = 1.000$. With only 4 matched conditions, these are descriptive inferential summaries over a deterministic toy classroom: the effect size is useful for scale, the permutation test is intentionally finite and seeded, and with 4 pairs it is underpowered to the point of being uninformative — $p = 1.000$ here means the test cannot distinguish this gap from sign-flip noise at all, which we state plainly rather than dressing as near-significance. The *sign* of the point estimate is the prediction the active-inference reading makes — a generative model conditioned on privileged beliefs (the teacher) should resolve more uncertainty per step than a variational posterior generating its own observations [Friston, 2013, Kirchhoff et al., 2018, Parr et al., 2020, Zhao et al., 2026, Jin et al., 2026] — and the measured mean gap is positive. We state plainly what the interval shows: at this sample size the confidence interval includes zero and the permutation test cannot reject the null, so the per-decision series demonstrates the direction of the effect and the honesty of the reporting machinery, not statistical significance. The thesis does not rest on this inferential summary: the correspondence is carried by the analytical identities, the executable models, and the deterministic mean entropy gap; this small-sample analysis is illustrative of honest reporting machinery, not confirmatory evidence. We frame these numbers as toy-classroom inferential summary, not a production-scale population claim: the confidence interval, effect size, and permutation p characterize sampling uncertainty within this toy classroom rather than asserting that the same advantage transfers to language-model distillation [Qwen Team, 2025, Lu and Thinking Machines Lab, 2025].

Simulation invariants merge into the analytical report after the pymdp harness runs (sec. 10). fig. 36 summarizes pass/fail status for both domains on the clean tree.

The replay matrix exposes deterministic rerun comparison as table data rather than prose. It contains 14 producer rows, uses explicit replay-or-fingerprint methods, and every row must match its saved artifact hash (`true`).

The `sensitivity` fragment binds the deterministic toy sweep to the canonical sheaf track. `output/data/sensitivity_sweep.json` contains 96 cells across toy parameters, planner labels, seeds, horizons, and graph topologies; the hydrated flag `true` is the only manuscript claim about coverage.

The companion `output/data/si_policy_grid.json` records measured policy-mode rows derived from `si_policy_comparison.json`, not a synthetic grid. Missing cells fail the artifact schema before they can become prose; the topology trace artifact contributes 4 deterministic topology traces.

The `uncertainty` fragment reports only normalized toy summaries. `output/data/uncertainty_summary.json` contains 21 belief and policy-posterior rows plus 3 finite entropy bins, and `true` is false if any posterior row fails to sum to one within the deterministic tolerance.

Policy uncertainty is recorded in generated policy artifacts rather than hand-entered into the manuscript. The posterior grid contributes 14 available posterior rows; the EFE values artifact reports availability-or-measured-fallback flag 1. The fragment is therefore a validation surface, not an empirical uncertainty claim.

The `benchmark` fragment adds a compact toy matrix over the Bernoulli, T-maze, and graph-world artifacts. `output/data/toy_benchmark_matrix.json` reports 3 model rows and `true` only when each row names an artifact, metric, and passing gate.

The matrix is scoped to deterministic study models. It is useful as a cross-track smoke test, not as a performance benchmark for biological or deployed systems.

The appendix `manuscript_staleness` row points to `output/reports/manuscript_staleness_report.json`. It checks 421 token

bindings after hydration, including late audit variables, and the pass flag is `true`.

This is the rendered-output side of the sheaf contract. Source fragments may contain hydration placeholders, but the public manuscript must not; the staleness report compares each token's generated value against the resolved markdown so stale counts are caught after composition, not only during source-file linting.

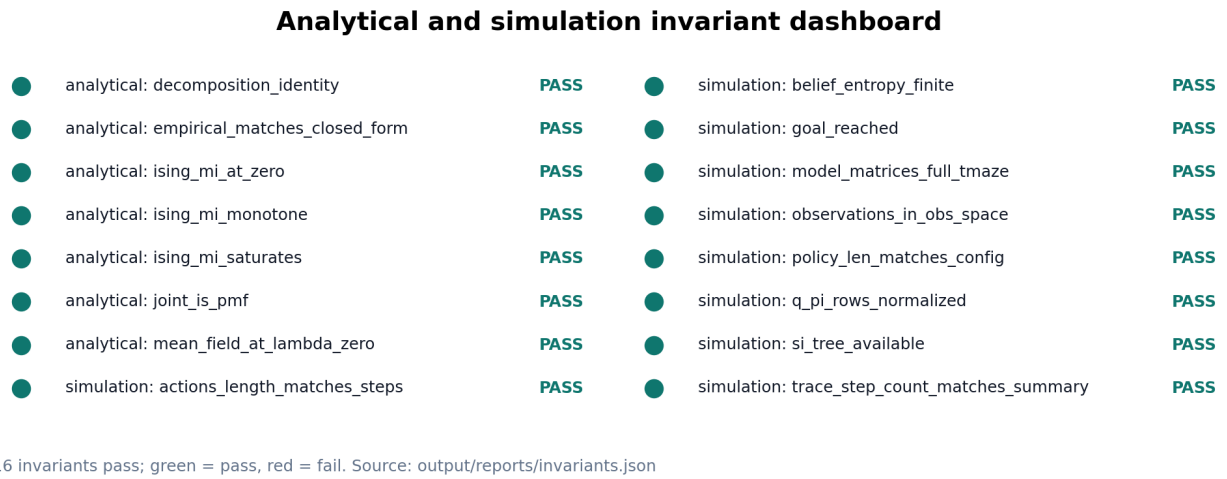


Figure 36: Invariant dashboard summarizing pass/fail status for every analytical and simulation check in the validation registry: 16 of 16 merged checks pass on the combined report. These invariants are the machine-enforced correctness conditions – conservation of probability mass, divergence non-negativity, free-energy bounds, and rollout consistency – that bind the toy on-policy-distillation claims to the active-inference math. An all-green dashboard means the registered analytical and simulation invariants pass for the generated toy artifacts.

The project's Lean boundary modules declare horizon and coupling witnesses. Build with `lake build` in `lean/`; fig. 14 summarizes proved versus deferred statements for this boundary fragment.

`sheaf-track:model_checking` binds `output/reports/model_checking_witnesses.json` and the Lean theorem inventories. The appendix claim is exactly 12 finite exhaustive witnesses with pass status `true`; Lean graph-world topology coverage is 4 generated topology ids with all-witnessed flag `true`.

`theorem_traceability_matrix.json` provides the appendix proof for theorem traceability: 22 linked rows with status `true`.

15.0.1 Appendix track: proof extraction

`proof_extraction` binds `output/data/proof_extraction_index.json` into the full sheaf appendix. Extracted theorems: 22. Constructive status: `true`.

The extraction index is intentionally modest: it records theorem names, statements, source files, leading tactics, and forbidden proof-token checks. That makes the Lean boundary inspectable without pretending that every proof term has been translated into a proof object. A row with a missing statement or forbidden token fails the formal interop gate and the canonical sheaf gate.

`output/data/proof_dependency_graph.json` adds the dependency view used by the appendix figure. It contributes 397 theorem-source, theorem-tactic, theorem-definition, and theorem-witness edges, with resolved edge status `true`; this is the artifact that keeps the theorem-traceability graph tied to generated Lean and model-checking rows.

15.0.2 State-space catalog track

The `state_space_catalog` track enumerates finite state spaces, action spaces, and policy counts for the deterministic toy models. The catalog artifact is `output/data/state_space_catalog.json`: it currently records 6 rows, with finite-space status `true`.

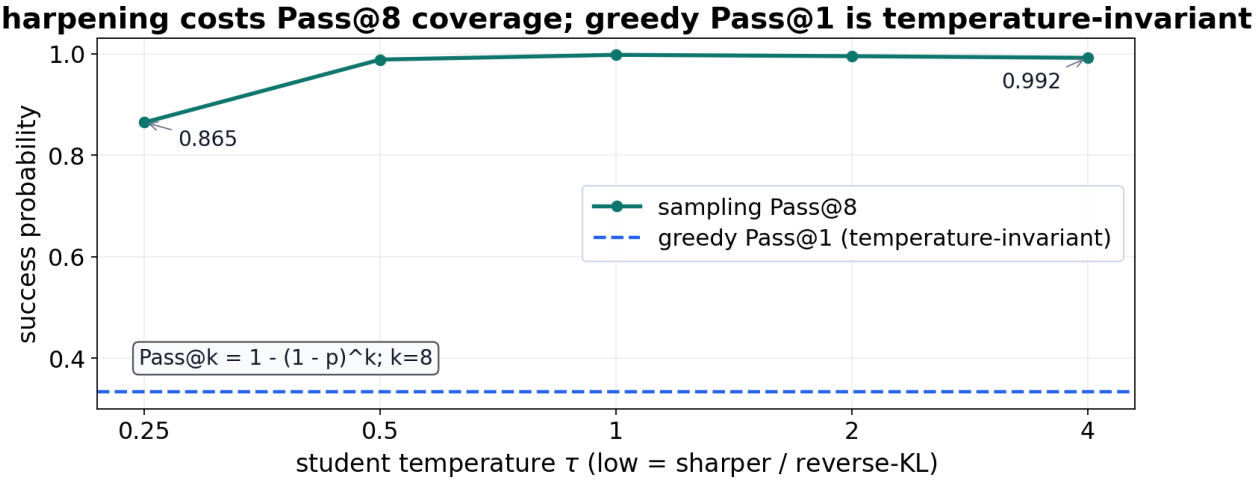
15.0.3 Causal ablation track

The `causal_ablation` track records deterministic toy ablations over finite preferences, likelihood-noise settings, and graph-topology perturbations. The matrix artifact is `output/data/causal_ablation_matrix.json`: it currently records 36 cells, with complete-grid status `true`.

GNN declarations: `gnn/bernoulli_toy.gnn.md` and `gnn/si_tmaze.gnn.md` [Smékal and Friedman, 2023]. fig. 10 and sec. 5 document ontology concordance for the Bernoulli toy; SI notation extends the same pattern under sec. 6.

15.0.4 Ontology bindings

- `belief_entropy` → **BeliefEntropy**
- `expected_free_energy` → **ExpectedFreeEnergy**
- `location` → **HiddenState**



Source: output/data/firstprinciples/diversity_demo.json

Figure 37: The diversity-collapse tradeoff of mode-seeking distillation, evaluated over a problem ensemble. Greedy Pass-at-1 (dashed, 0.333) is temperature-invariant; sampling Pass-at-k falls from 0.992 at the flattest temperature to 0.865 at the sharpest, because $Pass@k = 1-(1-p)^k$ for independent samples. Every curve is derived analytically from the declared temperature-sharpened ensemble (closed form, no sampling): this panel is an exact calculation over the toy problem ensemble, not an empirical measurement. Aggressive sharpening can raise single-answer commitment while lowering multi-sample coverage – the Pass-at-1-versus-Pass-at-k tension active inference frames as precision (inverse temperature), adjacent to the broader generation literature on objective- and decoding-induced diversity loss. Source: `output/data/firstprinciples/diversity_demo.json`.

- `observation` → **ObservationLikelihood**
- `policy` → **PolicyPosterior**
- `sheaf_track` → **SheafFragment**

Animation is an **extension** sheaf track backed by a deterministic GIF from `scripts/render_animation.py`. This appendix row documents the track binding only; default publication still uses static SI figures (sec. 10, fig. 24) while the GIF remains an auditable generated artifact.

The appendix `animation_delta` row points to `output/data/animation_frame_deltas.json`. The manifest records 3 adjacent-frame deltas, with `true` as the hydrated evidence that the GIF is trace-derived rather than a duplicated static frame.

15.0.5 Appendix track: release notes evidence

`release_notes` binds `output/reports/release_notes_evidence.json` into the full sheaf appendix. Rows: 3. Source-backed: `true`.

Release notes are treated as claims, not as informal changelog prose. Each row names a source artifact and a pass/deferred status, so the release note can say only what validation, bundle, or semantic artifacts support. The validator re-derives support from rows; flipping the summary bit without fixing a failed row still fails.

`output/reports/release_attestation.json` is the compact final view over the same boundary. It records 5 attestation rows for validation, release bundle hash, license audit, semantic certificate, and blocked-scope status, with all-attested flag `true`.

16 References

See `manuscript/references.bib` for bibliography entries cited in the composed sections.

References

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Rémi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. *arXiv preprint arXiv:1806.06920*, 2018. doi: 10.48550/arXiv.1806.06920. URL https://arxiv.org/abs/1806.06920.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2306.13649. URL https://arxiv.org/abs/2306.13649.

Miguel Aguilera, Beren Millidge, Alexander Tschantz, and Christopher L. Buckley. How particular is the physics of the free energy principle? *Physics of Life Reviews*, 40:24–50, 2022. doi: 10.1016/j.plrev.2021.11.001. URL https://doi.org/10.1016/j.plrev.2021.11.001.

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Chi Kit Cheung. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 700–710, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.58. URL https://aclanthology.org/2022.findings-acl.58/.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi: 10.1109/CVPR52729.2023.01499. URL https://arxiv.org/abs/2301.08243.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. doi: 10.48550/arXiv.1506.03099. URL https://papers.nips.cc/paper/5956-scheduled-sampling-for-sequence-prediction-with-recurrent-neural-networks.

Martin Biehl, Felix A. Pollock, and Ryota Kanai. A technical critique of some parts of the free energy principle. *Entropy*, 23(3):293, 2021. doi: 10.3390/e23030293. URL https://arxiv.org/abs/2001.06408.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024. doi: 10.48550/arXiv.2402.15391. URL https://arxiv.org/abs/2402.15391.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 535–541, 2006. doi: 10.1145/1150402.1150464. URL https://doi.org/10.1145/1150402.1150464.

Christopher L. Buckley, Chang Sub Kim, Simon McGregor, and Anil K. Seth. The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81:55–79, 2017. doi: 10.1016/j.jmp.2017.09.004. URL https://doi.org/10.1016/j.jmp.2017.09.004.

Yang Cai, Xiangyu Liu, Argyris Oikonomou, and Kaiqing Zhang. Provable partially observable reinforcement learning with privileged information. *arXiv preprint arXiv:2412.00985*, 2024. doi: 10.48550/arXiv.2412.00985. URL https://arxiv.org/abs/2412.00985.

Théophile Champion, Howard Bowman, Dimitrije Marković, and Marek Grès. Reframing the expected free energy: Four formulations and a unification. *arXiv preprint arXiv:2402.14460*, 2024. doi: 10.48550/arXiv.2402.14460. URL https://arxiv.org/abs/2402.14460.

Xianwei Chen, Shimin Zhang, and Jibin Wu. f-OPD: Stabilizing Long-Horizon On-Policy Distillation with Freshness-Aware Control. *arXiv preprint arXiv:2605.17862*, 2026. doi: 10.48550/arXiv.2605.17862. URL https://arxiv.org/abs/2605.17862.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, second edition, 1988. URL https://www.taylorfrancis.com/books/mono/10.4324/9780203771587/statistical-power-analysis-behavioral-sciences-jacob-cohen.

Louis Anthony Cox. Integrating fragmented risk knowledge: Sheaf theory for risk analysts. *Risk Analysis*, 46(3):e70206, 2026. doi: 10.1111/risa.70206. URL https://pubmed.ncbi.nlm.nih.gov/41742688/.

Justin Michael Curry. *Sheaves, Cosheaves and Applications*. PhD thesis, University of Pennsylvania, 2014. URL https://repository.upenn.edu/entities/publication/c391a10c-f2e5-40be-bc3f-e6f73a43ddfb.

Wojciech Marian Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant M. Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. Distilling policy distillation. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1331–1340. PMLR, 2019. doi: 10.48550/arXiv.1902.02186. URL https://proceedings.mlr.press/v89/czarnecki19a.html.

Lancelot Da Costa, Thomas Parr, Noor Sajid, Sebastijan Veselic, Victorita Neacsu, and Karl Friston. Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology*, 99:102447, 2020. doi: 10.1016/j.jmp.2020.102447. URL https://doi.org/10.1016/j.jmp.2020.102447.

Bert de Vries, Wouter Nuijten, Thijs van de Laar, Wouter Kouw, Sepideh Adamiat, Tim Nisslbeck, Mykola Lukashchuk, Hoang Minh Huu Nguyen, Marco Hidalgo Araya, Raphael Tresor, Thijs Jenneskens, Ivana Nikoloska, Raaja Ganapathy Subramanian, Bart van Erp, Dmitry Bagaev, and Albert Podusenko. Expected free energy-based planning as variational inference. *arXiv preprint arXiv:2504.14898*, 2025. doi: 10.48550/arXiv.2504.14898. URL https://arxiv.org/abs/2504.14898.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. doi: 10.48550/arXiv.2501.12948. URL https://arxiv.org/abs/2501.12948.

Matthew Fellows, Anuj Mahajan, Tim G. J. Rudner, and Shimon Whiteson. VIREL: A variational inference framework for reinforcement learning. In *Advances in Neural Information Processing Systems*, 2019. doi: 10.48550/arXiv.1811.01132. URL https://arxiv.org/abs/1811.01132.

Brendan Fong and David I. Spivak. *An Invitation to Applied Category Theory: Seven Sketches in Compositionality*. Cambridge University Press, 2019. doi: 10.1017/9781108668804. URL https://doi.org/10.1017/9781108668804.

Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. doi: 10.1038/nrn2787. URL https://www.nature.com/articles/nrn2787.

Karl Friston. Life as we know it. *Journal of the Royal Society Interface*, 10(86):20130475, 2013. doi: 10.1098/rsif.2013.0475. URL https://doi.org/10.1098/rsif.2013.0475.

Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1–3):70–87, 2006. doi: 10.1016/j.jphysparis.2006.10.001. URL https://doi.org/10.1016/j.jphysparis.2006.10.001.

Karl Friston, Lancelot Da Costa, Danijar Hafner, Casper Hesp, and Thomas Parr. Sophisticated inference. *Neural Computation*, 33 (3):713–763, 2021a. doi: 10.1162/neco_a_01351. URL https://doi.org/10.1162/neco_a_01351.

Karl J. Friston, Jean Daunizeau, and Stefan J. Kiebel. Reinforcement learning or active inference? *PLoS ONE*, 4(7):e6421, 2009. doi: 10.1371/journal.pone.0006421. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0006421.

Karl J. Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: A process theory. *Neural Computation*, 29(1):1–49, 2017a. doi: 10.1162/NECO_a_00912. URL https://doi.org/10.1162/NECO_a_00912.

Karl J. Friston, Min Lin, Christopher D. Frith, Giovanni Pezzulo, J. Allan Hobson, and Sasha Ondobaka. Active inference, curiosity and insight. *Neural Computation*, 29(10):2633–2683, 2017b. doi: 10.1162/neco_a_00999. URL https://doi.org/10.1162/neco_a_00999.

Karl J. Friston, Richard Rosch, Thomas Parr, Cathy Price, and Howard Bowman. Deep temporal models and active inference. *Neuroscience and Biobehavioral Reviews*, 90:486–501, 2018. doi: 10.1016/j.neubiorev.2018.04.004. URL https://doi.org/10.1016/j.neubiorev.2018.04.004.

Karl J. Friston, Lancelot Da Costa, and Thomas Parr. Some interesting observations on the free energy principle. *Entropy*, 23(8):1076, 2021b. doi: 10.3390/e23081076. URL https://doi.org/10.3390/e23081076.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models. In *International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2306.08543. URL https://arxiv.org/abs/2306.08543.

Anthony GX-Chen, Jatin Prakash, Jeff Guo, Rob Fergus, and Rajesh Ranganath. Kl-regularized reinforcement learning is designed to mode collapse. *arXiv preprint arXiv:2510.20817*, 2025. doi: 10.48550/arXiv.2510.20817. URL https://arxiv.org/abs/2510.20817.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, 2018. doi: 10.48550/arXiv.1809.01999. URL https://arxiv.org/abs/1809.01999.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 2018. doi: 10.48550/arXiv.1801.01290. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint*, 2023. doi: 10.48550/arXiv.2301.04104. URL https://arxiv.org/abs/2301.04104.

Zihao Han, Tiangang Zhang, Huaibin Wang, and Yilun Sun. Adaptive teacher exposure for self-distillation in LLM reasoning. *arXiv preprint arXiv:2605.11458*, 2026. doi: 10.48550/arXiv.2605.11458. URL https://arxiv.org/abs/2605.11458.

Tianxing He, Jingzhao Zhang, Zhiming Zhou, and James Glass. Exposure bias versus self-recovery: Are distortions really incremental for autoregressive text generation? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5087–5102, 2021. doi: 10.18653/v1/2021.emnlp-main.415. URL https://arxiv.org/abs/1905.10617.

Conor Heins, Beren Millidge, Daphne Demekas, Brennan Klein, Karl Friston, Iain D. Couzin, and Alexander Tschantz. pymdp: A python library for active inference in discrete state spaces. *Journal of Open Source Software*, 7(73):4098, 2022. doi: 10.21105/joss.04098. URL https://joss.theoj.org/papers/10.21105/joss.04098.

José Miguel Hernández-Lobato, Yingzhen Li, Mark Rowland, Daniel Hernández-Lobato, Thang Bui, and Richard E. Turner. Black-box α-divergence minimization. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, volume 48 of *Proceedings of Machine Learning Research*, pages 1511–1520. PMLR, 2016. doi: 10.48550/arXiv.1511.03243. URL https://proceedings.mlr.press/v48/hernandez-lobatob16.html.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. doi: 10.48550/arXiv.1503.02531. URL https://arxiv.org/abs/1503.02531.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. doi: 10.48550/arXiv.1904.09751. URL https://arxiv.org/abs/1904.09751.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026. doi: 10.48550/arXiv.2601.20802. URL https://arxiv.org/abs/2601.20802.

Ferenc Huszár. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*, 2015. doi: 10.48550/arXiv.1511.05101. URL https://arxiv.org/abs/1511.05101.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. *arXiv preprint arXiv:2601.07155*, 2026. doi: 10.48550/arXiv.2601.07155. URL https://arxiv.org/abs/2601.07155.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026. doi: 10.48550/arXiv.2603.07079. URL https://arxiv.org/abs/2603.07079.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999. doi: 10.1023/A:1007665907178. URL https://doi.org/10.1023/A:1007665907178.

Liyiming Ke, Sanjiban Choudhury, Matt Barnes, Wen Sun, Gilwoo Lee, and Siddhartha Srinivasa. Imitation learning as f-divergence minimization. In *Workshop on the Algorithmic Foundations of Robotics*, 2019. doi: 10.48550/arXiv.1905.12888. URL https://arxiv.org/abs/1905.12888.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Michael Kirchhoff, Thomas Parr, Ensor Palacios, Karl Friston, and Julian Kiverstein. The markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, 15(138):20170792, 2018. doi: 10.1098/rsif.2017.0792. URL https://doi.org/10.1098/rsif.2017.0792.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. doi: 10.48550/arXiv.2402.03898. URL https://arxiv.org/abs/2402.03898.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. Distillm-2: A contrastive approach boosts the distillation of llms. *arXiv preprint arXiv:2503.07067*, 2025. doi: 10.48550/arXiv.2503.07067. URL https://arxiv.org/abs/2503.07067.

Magnus Koudahl, Thijs van de Laar, and Bert de Vries. Realising synthetic active inference agents, part i: Epistemic objectives and graphical specification language. *arXiv preprint arXiv:2306.08014*, 2023. doi: 10.48550/arXiv.2306.08014. URL https://arxiv.org/abs/2306.08014.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-22/issue-1/On-Information-and-Sufficiency/10.1214/aoms/1177729694.full.

Aristotelis Lazaridis, Dylan Bates, Aman Sharma, Brian King, Vincent Lu, and Jack FitzGerald. EDGE-OPD: Internalizing privileged context with evidence guided on-policy distillation. *arXiv preprint arXiv:2605.23493*, 2026. doi: 10.48550/arXiv.2605.23493. URL https://arxiv.org/abs/2605.23493.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview preprint, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018. doi: 10.48550/arXiv.1805.00909. URL https://arxiv.org/abs/1805.00909.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan-ang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026. doi: 10.48550/arXiv.2604.13016. URL https://arxiv.org/abs/2604.13016.

Chris Yuhao Liu. Awesome on-policy distillation, 2026. URL https://zenodo.org/records/19411493.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, and Shu Wu. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026a. doi: 10.48550/arXiv.2605.21924. URL https://arxiv.org/abs/2605.21924.

Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When are teacher tokens reliable? position-weighted on-policy self-distillation for reasoning, 2026b. URL https://arxiv.org/abs/2605.21606.

Xinyu Liu, Darryl Cherian Jacob, Yang Zhou, Jindong Wang, and Pan He. Oisd: On-policy internal self-distillation of language models. *arXiv preprint arXiv:2605.29089*, 2026c. doi: 10.48550/arXiv.2605.29089. URL https://arxiv.org/abs/2605.29089.

Yifeng Liu, Shiyuan Zhang, Yifan Zhang, and Quanquan Gu. SDPG: Self-distilled policy gradient (reference implementation). https://github.com/lauyikfung/SDPG, 2026d. URL https://github.com/lauyikfung/SDPG.

Yifeng Liu, Shiyuan Zhang, Yifan Zhang, and Quanquan Gu. Self-distilled policy gradient. *arXiv preprint arXiv:2606.04036*, 2026e. doi: 10.48550/arXiv.2606.04036. URL https://arxiv.org/abs/2606.04036.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations (ICLR)*, 2016. doi: 10.48550/arXiv.1511.03643. URL https://arxiv.org/abs/1511.03643.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. URL https://thinkingmachines.ai/blog/on-policy-distillation/. Non-peer-reviewed industry research note (Connectionism series); cited only as external context, not as archival evidence.

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Zicheng Xu, Xiaotian Han, Tianyi Zhang, and Vladimir Braverman. Demystifying OPD: Length inflation and stabilization strategies for large language models. *arXiv preprint arXiv:2604.08527*, 2026. doi: 10.48550/arXiv.2604.08527. URL https://arxiv.org/abs/2604.08527.

Beren Millidge, Alexander Tschantz, Anil K. Seth, and Christopher L. Buckley. On the relationship between active inference and control as inference. In *Active Inference*, volume 1326 of *Communications in Computer and Information Science*, pages 3–11. Springer, 2020a. doi: 10.1007/978-3-030-64919-7_1. URL https://arxiv.org/abs/2006.12964.

Beren Millidge, Alexander Tschantz, Anil K. Seth, and Christopher L. Buckley. Reinforcement learning as iterative and amortised inference. *arXiv preprint arXiv:2006.10524*, 2020b. doi: 10.48550/arXiv.2006.10524. URL https://arxiv.org/abs/2006.10524.

Beren Millidge, Anil K. Seth, and Christopher L. Buckley. A mathematical walkthrough and discussion of the free energy principle. *arXiv preprint arXiv:2108.13343*, 2021a. doi: 10.48550/arXiv.2108.13343. URL https://arxiv.org/abs/2108.13343.

Beren Millidge, Alexander Tschantz, and Christopher L. Buckley. Whence the expected free energy? *Neural Computation*, 33(2): 447–482, 2021b. doi: 10.1162/neco_a_01354. URL https://doi.org/10.1162/neco_a_01354.

Brendan O'Donoghue, Ian Osband, and Catalin Ionescu. Making sense of reinforcement learning and probabilistic inference. In *International Conference on Learning Representations (ICLR)*, 2020. doi: 10.48550/arXiv.2001.00805. URL https://arxiv.org/abs/2001.00805.

Minjae Oh, Sangjun Song, Gyubin Choi, Yunho Choi, and Yohan Jo. Kl for a kl: On-policy distillation with control variate baseline. *arXiv preprint arXiv:2605.07865*, 2026. doi: 10.48550/arXiv.2605.07865. URL https://arxiv.org/abs/2605.07865.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. doi: 10.48550/arXiv.2203.02155. URL https://arxiv.org/abs/2203.02155.

Thomas Parr and Karl J. Friston. Generalised free energy and active inference. *Biological Cybernetics*, 113(5–6):495–513, 2019. doi: 10.1007/s00422-019-00805-w. URL https://doi.org/10.1007/s00422-019-00805-w.

Thomas Parr, Lancelot Da Costa, and Karl J. Friston. Markov blankets, information geometry and stochastic thermodynamics. *Philosophical Transactions of the Royal Society A*, 378(2164):20190159, 2020. doi: 10.1098/rsta.2019.0159. URL https://doi.org/10.1098/rsta.2019.0159.

Thomas Parr, Giovanni Pezzulo, and Karl J. Friston. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press, 2022. doi: 10.7551/mitpress/12441.001.0001. URL https://mitpress.mit.edu/9780262045353/active-inference/.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*, 2026a. doi: 10.48550/arXiv.2602.04942. URL https://arxiv.org/abs/2602.04942.

Emiliano Penaloza, Dheeraj Vattikonda, Siddarth Venkatraman, and Massimo Caccia. Understanding self-distillation and privileged information distillation. Interactive tutorial, https://emilianopp.github.io/Privileged-Information-Distillation-and-Self-Distillation/, 2026b. URL https://emilianopp.github.io/Privileged-Information-Distillation-and-Self-Distillation/.

Steven Phillips. Sheaving: A universal construction for semantic compositionality. *Philosophical Transactions of the Royal Society B*, 375(1791):20190303, 2020. doi: 10.1098/rstb.2019.0303. URL https://doi.org/10.1098/rstb.2019.0303.

Dean A. Pomerleau. ALVINN: An autonomous land vehicle in a neural network. In *Advances in Neural Information Processing Systems*, volume 1, pages 305–313, 1989. URL https://proceedings.neurips.cc/paper/1988/hash/812b4ba287f5ee0bc9d43bbf5bbe87fb-Abstract.html.

Andrea Pozzi, Alessandro Incremona, Daniele Tessera, and Daniele Toti. Mitigating exposure bias in large language model distillation: An imitation learning approach. *Neural Computing and Applications*, 37:12013–12029, 2025. doi: 10.1007/s00521-025-11162-0. URL https://doi.org/10.1007/s00521-025-11162-0.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023. doi: 10.48550/arXiv.2305.18290. URL https://arxiv.org/abs/2305.18290.

Miguel Moura Ramos, Duarte M. Alves, and André F. T. Martins. Combining on-policy optimization and distillation for long-context reasoning in large language models. *arXiv preprint arXiv:2605.12227*, 2026. doi: 10.48550/arXiv.2605.12227. URL https://arxiv.org/abs/2605.12227.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*, 2016. doi: 10.48550/arXiv.1511.06732. URL https://arxiv.org/abs/1511.06732.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. doi: 10.1038/4580. URL https://doi.org/10.1038/4580.

Michael Robinson. *Topological Signal Processing*. Springer, 2014. doi: 10.1007/978-3-642-36104-3. URL https://link.springer.com/book/10.1007/978-3-642-36104-3.

Michael Robinson. Sheaves are the canonical data structure for sensor integration. *Information Fusion*, 36:208–224, 2017. doi: 10.1016/j.inffus.2016.12.002. URL https://doi.org/10.1016/j.inffus.2016.12.002.

Dhruv Rohatgi, Adam Block, Audrey Huang, Akshay Krishnamurthy, and Dylan J. Foster. Computational-statistical tradeoffs at the next-token prediction barrier: Autoregressive and imitation learning under misspecification. *arXiv preprint arXiv:2502.12465*, 2025. doi: 10.48550/arXiv.2502.12465. URL https://arxiv.org/abs/2502.12465.

Daniel Rosiak. *Sheaf Theory through Examples*. MIT Press, 2022. doi: 10.7551/mitpress/12581.001.0001. URL https://doi.org/10.7551/mitpress/12581.001.0001.

Stephane Ross and J. Andrew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 661–668. PMLR, 2010. URL https://proceedings.mlr.press/v9/ross10a.html.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 627–635. PMLR, 2011. doi: 10.48550/arXiv.1011.0686. URL https://proceedings.mlr.press/v15/ross11a.html.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2016. doi: 10.48550/arXiv.1511.06295. URL https://arxiv.org/abs/1511.06295.

Noor Sajid, Philip J. Ball, Thomas Parr, and Karl J. Friston. Active inference: Demystified and compared. *Neural Computation*, 33 (3):674–712, 2021a. doi: 10.1162/neco_a_01357. URL https://doi.org/10.1162/neco_a_01357.

Noor Sajid, Lancelot Da Costa, Thomas Parr, and Karl J. Friston. Active inference, bayesian optimal design, and expected utility. *arXiv preprint arXiv:2110.04074*, 2021b. doi: 10.48550/arXiv.2110.04074. URL https://arxiv.org/abs/2110.04074.

Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten simple rules for reproducible computational research. *PLoS Computational Biology*, 9(10):e1003285, 2013. doi: 10.1371/journal.pcbi.1003285. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003285.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, Chess and Shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020. doi: 10.1038/s41586-020-03051-4. URL https://arxiv.org/abs/1911.08265.

Michal Sharoni and Sivan Sabato. On the capacity limits of privileged erm. *Proceedings of the Twenty Sixth International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR 206*, pages 523–534, 2023. doi: 10.48550/arXiv.2303.02658. URL https://arxiv.org/abs/2303.02658.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026. doi: 10.48550/arXiv.2601.19897. URL https://arxiv.org/abs/2601.19897.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. doi: 10.1016/S0378-3758(00)00115-4. URL https://doi.org/10.1016/S0378-3758(00)00115-4.

Aman Shrivastava, Yanjun Qi, and Vicente Ordonez. Estimating and maximizing mutual information for knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. doi: 10.48550/arXiv.2110.15946. URL https://arxiv.org/abs/2110.15946.

Jakub Smékal and Daniel Ari Friedman. Generalized notation notation for active inference models, 2023. URL https://zenodo.org/records/7803328.

Ryan Smith, Karl J. Friston, and Christopher J. Whyte. A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*, 107:102632, 2022. doi: 10.1016/j.jmp.2021.102632. URL https://doi.org/10.1016/j.jmp.2021.102632.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. *arXiv preprint arXiv:2209.15189*, 2022. doi: 10.48550/arXiv.2209.15189. URL https://arxiv.org/abs/2209.15189.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026. doi: 10.48550/arXiv.2604.00626. URL https://arxiv.org/abs/2604.00626.

Alberto Speranzon, David I. Spivak, and Srivatsan Varadarajan. Abstraction, composition and contracts: A sheaf theoretic approach. *arXiv preprint arXiv:1802.03080*, 2018. doi: 10.48550/arXiv.1802.03080. URL https://arxiv.org/abs/1802.03080.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, and Andrew Gordon Wilson. Does knowledge distillation really work? In *Advances in Neural Information Processing Systems*, volume 34, 2021. doi: 10.48550/arXiv.2106.05945. URL https://proceedings.neurips.cc/paper/2021/hash/376c6b9ff3bedbbea56751a84fffc10c-Abstract.html.

Wen Sun, Arun Venkatraman, Geoffrey J. Gordon, Byron Boots, and J. Andrew Bagnell. Deeply aggrevated: Differentiable imitation learning for sequential prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3309–3318. PMLR, 2017. doi: 10.48550/arXiv.1703.01030. URL https://proceedings.mlr.press/v70/sun17a.html.

Kanghui Tian, Siyuan Liu, Ziang Yan, Sheng Xia, Shuai Dong, and Yi Wang. ViCuR: Visual cues as recoverable privilege for multimodal on-policy distillation. *arXiv preprint arXiv:2606.05718*, 2026. doi: 10.48550/arXiv.2606.05718. URL https://arxiv.org/abs/2606.05718.

Emanuel Todorov. General duality between optimal control and estimation. In *Proceedings of the 47th IEEE Conference on Decision and Control*, pages 4286–4292, 2008. doi: 10.1109/CDC.2008.4739438. URL https://doi.org/10.1109/CDC.2008.4739438.

Marc Toussaint. Robot trajectory optimization using approximate inference. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1049–1056, 2009. doi: 10.1145/1553374.1553508. URL https://doi.org/10.1145/1553374.1553508.

Alexander Tschantz, Manuel Baltieri, Anil K. Seth, and Christopher L. Buckley. Scaling active inference. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020a. doi: 10.1109/IJCNN48605.2020.9207382. URL https://arxiv.org/abs/1911.10601.

Alexander Tschantz, Beren Millidge, Anil K. Seth, and Christopher L. Buckley. Reinforcement learning through active inference. *arXiv preprint arXiv:2002.12636*, 2020b. doi: 10.48550/arXiv.2002.12636. URL https://arxiv.org/abs/2002.12636.

Jesse van Oostrum, Carlotta Langer, and Nihat Ay. A concise mathematical description of active inference in discrete time. *arXiv preprint arXiv:2406.07726*, 2024. doi: 10.48550/arXiv.2406.07726. URL https://arxiv.org/abs/2406.07726.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009. doi: 10.1016/j.neunet.2009.06.042. URL https://doi.org/10.1016/j.neunet.2009.06.042.

Mark D. Wilkinson et al. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016. doi: 10.1038/sdata.2016.18. URL https://www.nature.com/articles/sdata201618.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking kullback-leibler divergence in knowledge distillation for large language models. *arXiv preprint arXiv:2404.02657*, 2024. doi: 10.48550/arXiv.2404.02657. URL https://arxiv.org/abs/2404.02657.

Yecheng Wu, Song Han, and Hai Cai. Lightning opd: Efficient post-training for large reasoning models with offline on-policy distillation, 2026. URL https://arxiv.org/abs/2604.13010.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation. *arXiv preprint arXiv:2606.01249*, 2026. doi: 10.48550/arXiv.2606.01249. URL https://arxiv.org/abs/2606.01249.

Wenda Xu, Rujun Han, Zifeng Wang, Long T. Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. *arXiv preprint arXiv:2410.11325*, 2024. doi: 10.48550/arXiv.2410.11325. URL https://arxiv.org/abs/2410.11325.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1028–1043, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.58. URL https://aclanthology.org/2024.acl-long.58/.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*, 2025. doi: 10.48550/arXiv.2511.10643. URL https://arxiv.org/abs/2511.10643.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026. doi: 10.48550/arXiv.2602.12275. URL https://arxiv.org/abs/2602.12275.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning. *arXiv preprint arXiv:2203.14465*, 2022. doi: 10.48550/arXiv.2203.14465. URL https://arxiv.org/abs/2203.14465.

Xinsen Zhang, Zhenkai Ding, Tianjun Pan, Run Yang, Chun Kang, Xue Xiong, and Jingnan Gu. OPSDL: On-policy self-distillation for long-context language models. *arXiv preprint arXiv:2604.17535*, 2026. doi: 10.48550/arXiv.2604.17535. URL https://arxiv.org/abs/2604.17535.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026. doi: 10.48550/arXiv.2601.18734. URL https://arxiv.org/abs/2601.18734.

Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Houcheng Jiang, Xiang Wang, and Junfeng Fang. SOD: Step-wise on-policy distillation for small language model agents. *arXiv preprint arXiv:2605.07725*, 2026. doi: 10.48550/arXiv.2605.07725. URL https://arxiv.org/abs/2605.07725.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes. *arXiv preprint arXiv:2605.11182*, 2026a. doi: 10.48550/arXiv.2605.11182. URL https://arxiv.org/abs/2605.11182.

Wenhong Zhu, Ruobing Xie, Rui Wang, and Pengfei Liu. Hybrid policy distillation for llms. *arXiv preprint arXiv:2604.20244*, 2026b. doi: 10.48550/arXiv.2604.20244. URL https://arxiv.org/abs/2604.20244.

Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pages 1433–1438, 2008. URL https://aaai.org/papers/022-maximum-entropy-inverse-reinforcement-learning/.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. doi: 10.48550/arXiv.1909.08593. URL https://arxiv.org/abs/1909.08593.

END OF TRANSMISSION

Release: v1.0.2 · DOI 10.5281/zenodo.20747834 · SHA-256 8d70985fca93… · pairing complete

Figure 38: Integrity QR strip

Prior: v1.0.0 · 10.5281/zenodo.20747834 · db0f2e4f… · v1.0.1 · 10.5281/zenodo.20748663 · 4f7040bc…