ENIGMA: The Geometry of Reasoning and Alignment in Large-Language Models

Gareth Seneque

Lap-Hang Ho

Nafise Erfanian Saeedi

Jeffrey Molendijk

Ariel Kuperman

Tim Elson

Australian Broadcasting Corporation

October 13, 2025

These facts are, like all facts, not necessary but of a merely empirical certainty; they are hypotheses; one may therefore inquire into their probability.

— *Bernard Riemann (1854)*

Geometry is the science of correct reasoning on incorrect figures.

— *George Pólya,* How to Solve It *(1945)*

Abstract

We present *Entropic Mutual-Information Geometry Large-Language Model Alignment* (ENIGMA), a novel approach to Large-Language Model (LLM) training that jointly improves reasoning, alignment and robustness by treating an organisation's policies/principles as directions to move on a model's information manifold. Our single-loop trainer combines *Group-Relative Policy Optimisation* (GRPO), an on-policy, critic-free RL method with Chain-of-Thought (CoT)-format only rewards; a *Self-Supervised Alignment with Mutual Information* (SAMI)-style symmetric InfoNCE auxiliary; and an entropic Sinkhorn optimal-transport regulariser on hidden-state distributions to bound geometry drift. We also introduce infoNCE metrics that specialise to a standard MI lower bound under matched negatives to measure how strongly a model's CoT encodes these policies. These metrics include a Sufficiency Index (SI) that enables the selection and creation of principles that maximise downstream performance prior to training. In our experiments using small (1B) LLMs, high-SI principles predict steadier training dynamics and improved benchmark performance over GRPO ablations. Our information-geometry analysis of trained models validates desirable structural change in the manifold. These results support our hypothesis that reasoning, alignment, and robustness are projections of a single information-geometric objective, and that models trained using ENIGMA demonstrate *principled reasoning* without the use of a reward model, offering a path to trusted capability.

1

Contents

1 Introduction

Authors' note: no editorial position is offered or claimed in this publication. All LLM prompts and synthetic data examples are provided to support replication of our work and do not reflect optimisation for production use or reflect an editorial position.

1.1 Overview

In our prior work on reasoning & alignment, *ABC Align* [42] we developed a post-training framework suitable for organisations who aim to leverage the capabilities of LLMs while making use of their data and preserving provider independence, proxies for organisational obligations to innovation and independence. To do this, we leveraged open research and techniques including Orca [39], Less Is More for Alignment (LIMA) [40], and Odds Ratio Preference Optimization (ORPO) [41] to make use of proprietary data (synthetic reasoning traces constructed from ABC content, conditioned on our AI Principles) to elicit improved performance on both alignment and reasoning benchmarks, avoiding the so-called alignment tax. These methods were effective but fundamentally post-hoc: they relied on in-context prompts or preference-style fine-tuning and offered limited control over how principles reshape the model's internal geometry.

ENIGMA departs from that setting. We treat written principles as *directions of motion on a model's information manifold*, and we design both the training signal and the measurements to act directly on that geometry. In doing so we aim to elicit *principled reasoning*: chains of thought that encode stated principles, without an external reward model and with explicit, reportable evidence that those principles are encoded.

Recent progress in "reasoning LLMs" largely couples on-policy RL with chain-of-thought (CoT) sampling, yielding strong gains where answer correctness can be verified (math, code). In subjective domains, where correctness encompasses normative definitions of truthfulness, alignment to preferences etc., correct-answer supervision is costly and fragile, and even visible reasoning traces can be unfaithful. Industry standard tooling also lacks the built-in capabilities to measure interventions that target information geometry objectives.

To address these limitations, ENIGMA starts from the premise that *neural networks are geometric entities*.

1.2 ENIGMA

In this paper, we introduce a single-loop training method that adapts Group-Relative Policy Optimisation (GRPO) [8], Self-Supervised Alignment with Mutual Information (SAMI) [9], and Sinkhorn divergence to implement Optimal Transport (OT) [10, 11] and develop quantitative measures to evaluate constitutional principles both prior to training and their impact on training itself. Our results provide evidence that we elicit *principled reasoning* from LLMs using open and proprietary data without the cost and complexity of human preference data collection or correctness-verified question/answer datasets for general domains. Our use of 'principled' reflects that our metrics provide lower-bound evidence that completions encode the constitutional principles. For our experiments, we have used the ABC's Editorial Policies [2] as 'positive' constitutions, a source of human-written truth, and two sets of LLM-generated 'negative' constitutions. We have developed a method of quantifying the constitution's impacts on training dynamics that predicts

downstream task performance. This method produces a novel Sufficiency Index (SI) that is generally applicable to any set of paired constitutions.

Our intervention is thus fundamental, going beyond standard preference optimisation techniques or alignment methods that involve exploiting the properties of In-Context Learning (ICL) via natural-language 'prompts' without updating the model's parameters, These methods are effective for a variety of use-cases (outlined in our work on *ABC Align* [4], e.g., generating content metadata or multi-turn conversation), but are ultimately brittle and amenable to input distribution perturbation (see evaluation discussions in recent surveys/best-practice notes [44]).

We necessarily limit the scope of our experiments to small LLMs given the setting in which this work was conducted, that of a public-service media organisation (see our prior work on *ABC Align* for additional organisational context [4] and our 'Limitations' section for additional detail). Our methods and results however demonstrate that ENIGMA ultimately enables organisational collaboration on defining principles and standards for any use-case where a high degree of alignment is required, and provides quantitative measures that relate those principles to model behaviour and outputs.

1.3 Key contributions

We make several contributions that connect Constitutional AI evaluation, verification-free RL, and MI/OT-based LLM alignment techniques.

- **Unified information-geometric training objective (ENIGMA).**
  We formulate *reasoning, alignment,* and *adversarial robustness* as a single optimization problem on the product space of distributions over next-token logits and sequence-level hidden representations by coupling on-policy GRPO [8] with Constitutional-principle shaping [9] and a Sinkhorn divergence penalty [10, 11], all in one loop. This bridges online RL with principle-guided supervision and geometry-aware regularisation. We see benchmark improvements for GPQA (main) +6.92pts/50.81%; and TruthfulQA [12, 13] +12.11pts/31.81% (absolute/relative scores)

- **Reportable MI lower bound for "principle encoding."**
  We introduce row/column 'clean' InfoNCE metrics [18], a computationally light alternative to the symmetric InfoNCE of SAMI [9]. The metric is 'clean' as it is calculated without entropy gating and with uniform shadow principles. It yields a per-step lower bound on how strongly completions encode a targeted principle. This gives a falsifiable, quantised proxy for "constitutional adherence" that can be tracked like loss/accuracy (note: we address known bias/variance pathologies at high MI by design of the metric.)

- **Principle curation by *effective MI*.**
  We propose quantitative sufficiency signals and a composite metric, a Sufficiency Index (SI) to *rewrite/select* constitutional principles, producing a set that empirically yields better training dynamics and downstream task performance. Constitutional AI established principle-guided training [6]; we add principle selection via MI diagnostics [18, 23].

- **Interpretability-friendly geometry probes.**
  We track Bhattacharyya angle, Hellinger, and JS divergence (metrics background in [45, 46, 50]; Equations (EqC-18) to (EqC-20)) between last-token distributions and representation-level

probes (Fréchet distance [47]; Equation (EqC-21), effective rank/participation ratio [48, 49]; Equation (EqC-22)) to connect manifold movement with task gains.

- **CoT-only & verifier-free reward 'tie-breaker' with GRPO.**
  We show that format-only rewards (XML tags for reasoning/answer) plus MI-based tie-breaking suffice to move reasoning quality and calibration, using on-policy GRPO [8] to avoid stale off-policy artifacts without the use of an external reward model. Our tie-breaker reward makes principle encoding actionable for GRPO without overwhelming task reward and allows scaling of our methods to larger models where format compliance is high from pre- and post-training.

- **Principled, *parameter-efficient* RL post-training.**
  As in our prior work, we leverage LoRA for memory efficiency, while enabling frequent MI/OT/geometry measurements. This approach is both cost- and compute-efficient.

2 Related Work

This section situates ENIGMA relative to three threads in post-training for LLMs. First, policy-gradient alignment with trust-region control: TRPO [14] and PPO [15] interpret the KL-to-reference constraint as a local step-size control that approximates natural-gradient motion on the Fisher–Rao manifold of next-token distributions; GRPO replaces a trained critic with a group baseline well-suited to multi-sample CoT pipelines [8].[1]

Second, mutual-information–based alignment: contrastive learning (InfoNCE) provides a variational lower bound on mutual information [18], and recent work (SAMI) uses conditional MI $I(Y; C \mid X)$ to teach models to follow written principles ("constitutions") without preference labels or demonstrations [9].

Third, optimal-transport regularization: entropic OT (Sinkhorn) yields a smooth, scalable divergence between empirical measures with well-understood convexity and convergence-in-law properties [10, 11]. Recent alignment work casts distributional preference alignment as an OT problem (AOT) that enforces relaxed first-order stochastic dominance over reward distributions [19].

We also cover complementary studies on process supervision (PRMs/CoT) and constitution selection, then gather the threads in a way that points to the theoretical basis of our work, outlined in the following section.

Our aim here is not an exhaustive survey but to surface design choices that motivated ENIGMA's single-loop combination of GRPO (policy control) [8], SAMI (principled process alignment in representation space) [9], and a lightweight Sinkhorn penalty on hidden states (distributional drift control) [10, 11].

2.1 Group Relative Policy Optimization (GRPO)

GRPO, introduced in *DeepSeekMath* [8], replaces a learned value critic with a group baseline computed from multiple sampled completions per prompt, while retaining a KL-to-reference trust

---

[1]Not to be confused with Group Robust Preference Optimisation, an unrelated "GRPO" acronym in preference optimization [20])

region (PPO-style). This removes the critic's memory/compute overhead and matches grouped sampling used in reasoning pipelines.

KL-constrained policy optimization (TRPO/PPO) can be interpreted as approximate natural-gradient steps on the Fisher information manifold [14, 15]; the trust region respects local distributional curvature. Related work defines Wasserstein natural gradients by pulling back the W2 metric to parameter space, motivating links between KL-style and OT-style geometries [16, 17]. For ENIGMA, we utilise PPO-style ratio clipping (DR-GRPO) to provide local step-control.[2]

2.2 Self-Supervised Alignment with Mutual Information (SAMI)

SAMI fine-tunes an LLM to maximize conditional mutual information $I(Y; C \mid X)$ between responses Y and a constitution C given prompts X, requiring neither preference labels nor demonstrations [9]. InfoNCE lower bounds MI and supplies a practical contrastive objective for such alignment [18]; see also analyses and caveats around MI maximization [23]. SAMI complements Constitutional AI (CAI), which uses a written constitution to produce AI feedback and then runs supervised/RL phases [6].

2.3 Optimal Transport (OT) regularisation

Optimal Transport (OT) offers a principled way to compare probability distributions by computing the minimal "work" to morph one into another. The resulting 2-Wasserstein metric equips the space of probability measures with a geometry that behaves Riemannian in the sense of Otto's calculus [23], so distributions can be connected by geodesics and even interpreted through gradient-flow dynamics. In practice, entropic regularisation leads to the Sinkhorn formulation, which is an efficient, differentiable approximation whose smoothness/positivity properties make it well-suited to modern deep-learning pipelines [10, 11]. Beyond serving as a geometry-aware distance, OT has become a useful regulariser for aligning model behaviours. AOT casts distributional preference alignment as a 1-D OT problem that enforces (relaxed) first-order stochastic dominance between reward distributions and reports strong results among 7B models [12]. In a vision-language adaptation, Prompt-OT uses an OT penalty to preserve feature-distribution structure during prompt tuning [12]. In this work we adopt the same spirit: a small entropic-OT term keeps representations within a shared W2 ball, limiting drift while leaving GRPO's low-variance updates and SAMI's principle-consistent directions intact [10, 11, 19].

2.4 Chain of Thought, Process Supervision and Monitoring

While Chain-of-Thought (CoT) prompting improves measured reasoning on benchmarks, its safety value depends on whether intermediate traces are *faithful* to the model's computations. Process supervision and process-reward models (PRMs) directly score intermediate steps and have been shown to outperform outcome-only supervision on math reasoning and to enable step-level audits (e.g., PRM800K; Let's Verify Step by Step) [4]. Recent work finds that reasoning models may rationalise or hide internal computations [21], underscoring the need for monitors that evaluate

---

[2]Note: "GRPO" is also used for Group Robust Preference Optimization in reward-free RLHF; to avoid confusion, we use GRPO (policy) for DeepSeekMath's algorithm [8] and GRPO (group-robust) for the unrelated worst-group PO method [55].

process rather than only outcomes (see also work on CoT monitorability [22]). Additionally, there is a growing literature on 'reasoning in latent space' [3], which may further limit interpretability and monitoring efforts. ENIGMA's CoT-format–only rewards and MI reward tiebreaker, coupled with SAMI-based shaping and loss, provide a reference- and verifier-free means of process supervision, where the information contained in a set of principles is encoded into the CoT process itself, which is produced as part of the model's completion.

2.5 Constitution Evaluation and Principled Reasoning

Existing approaches to deploying and adapting Constitutional AI (CAI) utilise human-written and synthetic principles but offer few quantitative tools for selecting or curating them toward a target metric that guides their suitability for the training process itself [6]. We address this gap by defining **SI**. For a given set of positive/negative constitutions or principles, we measure predictive information via ΔNLL, associative information via InfoNCE-style MI bounds [18, 23], and separation (ΔNLL AUC). We also track information-geometry observables (Bhattacharyya angle, Hellinger, effective rank) [22, 25]. to diagnose manifold changes and relate them to changes in task performance.

2.6 Towards Operational Guarantees for Large-Language Model Alignment

As LLMs transition from chat assistants or data processing tools to instruments that facilitate the automation of decision-making, international AI policy and standards guidance emphasises documented risk management, adversarial testing, and transparency [56, 57, 58]. ENIGMA supplies two complementary measures towards guarantees of safe, reliable capability: principle encoding via the MI lower bound discussed above (subject to known MI bound trade-offs [23]) and distributional alignment via entropic Sinkhorn OT that aligns preference/reward distributions rather than isolated samples, an approach recently shown to enforce stochastic-dominance-style constraints in LLM alignment [19].

Our implementation includes configuration parameters to stabilise this optimisation with Fisher-metric preconditioning (natural-gradient) and entropy control, standard tools that reduce mode collapse and training pathologies in policy-gradient style fine-tuning. We did not observe gradient conflicts during training, however for different settings and model capacities, the literature on multi-task learning offers solutions to this problem [59].

2.7 An information-geometric view of LLM post-training

The threads above admit a common geometric reading that will structure the next section. Policy updates live on the Fisher–Rao manifold of categorical token distributions; TRPO/PPO give a trust-region/natural-gradient view [14, 15]. Contrastive MI acts on ℓ_2-normalised representations; InfoNCE tightens a lower bound on $I(Y; C \mid X)$ [18] through the row/column objectives in Equations (EqC-06) to (EqC-08), moving paired (response, principle) embeddings along spherical geodesics (cf. hyperspherical geometry perspectives [26]). Sinkhorn OT treats per-sequence hidden states as empirical measures and penalizes coherent mass transport away from a reference via Equations (EqC-13) to (EqC-16).

We do not assume a formal product-manifold geodesic. Instead, we use this trio as a composite control system: local policy steps (Fisher–Rao) + semantic alignment in representation space

(spherical/InfoNCE) + global drift control over hidden-state distributions (Wasserstein/Sinkhorn) [15, 16, 19, 23].[3]

3 Information Geometry and Alignment

The Fisher–Rao view (TRPO/PPO/GRPO) endows token distributions with a local Riemannian structure; ratio-clipping (and KL, which we do not employ here) restricts step size in that geometry. The contrastive view (InfoNCE) acts on normalised representations, pulling completions toward their governing principles and away from negatives, thereby injecting a positive principle-aligned semantic direction, useful for domains where correctness is hard to verify. The Wasserstein view (entropic Sinkhorn) treats batches of hidden states as empirical measures and penalizes mass transport away from the reference, serving as a global guardrail against drift that appears small in KL but semantically large in representation space.

3.1 Policy space as a Fisher–Rao manifold

Consistent with existing literature, the family of categorical token distributions forms a statistical manifold with the Fisher information as its Riemannian metric; locally, the KL divergence between nearby policies equals ½ of the squared Fisher–Rao distance up to higher-order terms [50, 51]. This underpins natural-gradient methods' reparameterisation invariance and trust-region interpretations (Equation (EqC-05)).

TRPO directly constrains a per-update KL trust region (Equation (EqC-04)), yielding monotonic-improvement guarantees under its approximations; PPO [5] enforces an approximate trust region through clipped ratios and/or KL penalties via the surrogate in Equation (EqC-03); GRPO removes the critic and uses a group baseline for advantages (Equation (EqC-02)) while keeping the same KL-to-reference control and is well-suited to CoT-style grouped sampling. As noted, we do not use KL for our implementation[4], in line with recent GRPO best-practices instead we use PPO-style ratio clipping (DR-GRPO) to provide local step-control and instead employ Sinkhorn divergence for regularisation.

3.2 Contrastive alignment on a (hyper)sphere

When embeddings are ℓ_2 -normalised, they live on a unit hypersphere; contrastive objectives such as InfoNCE maximize a lower bound on mutual information by increasing the similarity of matched pairs relative to negatives [18] via the objectives in Equations (EqC-06) and (EqC-07). Hyperspherical models and geodesic perspectives motivate our geometric reading [26].

SAMI operationalises this view at the principle level: it maximizes the conditional MI $I(Y; C \mid X)$ between responses and a constitution-given prompt, without preference labels. In our experiments, that constitution foregrounds first-principles reasoning and factuality, expressed through the editorial standards of a PSM, biasing the model toward provable aligned chains-of-thought even without answer-correctness verification.

---

[3]See also Amari's information-geometry framing [24].

[4]Though we conducted many experiments with both KL and OT regularisation inside the same training loop, including an active KL controller, we observed a negative impact on training stability.

Because $I(Y; C \mid X)$ equals an expected KL between the joint $p(y, c \mid x)$ and its factorisation $p(y \mid x)p(c \mid x)$, the SAMI loss/reward shaper in Equations (EqC-08) and (EqC-12) shortens geodesics on the statistical manifold in directions predictable from the constitution, a semantic directional constraint that substitutes for the KL step-size constraint on the policy [59]. This is the theoretical lens which, combined with recent GRPO best-practices and observations from our early experiments, motivates our use of OT instead of KL regularisation.

3.3   Sinkhorn OT and the Wasserstein geometry of hidden-state measures

Sequences of hidden states can be treated as empirical measures. The squared 2-Wasserstein distance in Equation (EqC-13) measures geodesic distance in the space of probability measures (the Otto metric). We adopt Sinkhorn divergences, entropically-regularised OT with debiasing, which are positive, convex, and metrise convergence in law while remaining smooth and scalable for backprop [10, 11]; Equations (EqC-14) and (EqC-15) capture these quantities.

ENIGMA computes a Sinkhorn divergence between empirical measures of hidden states from the current policy and a frozen reference, penalizing shifts that look small in KL but correspond to coherent mass transport (e.g., moving probability mass between modes), acting as a geometry-aware tether [28] through the regulariser in Equation (EqC-16).

3.4   A product-manifold-inspired trust region

For each ENIGMA training step, GRPO supplies a local KL trust region for CoT-only rewards, MI supplies a semantic direction consistent with a first-principles/factuality constitution, and Sinkhorn OT provides a global geometry-aware constraint (via regularisation). While we use "product-manifold" language for intuition, this is an algorithmic composite rather than a formal geodesic in Equation (EqC-17); guarantees are inherited from the constituent terms (e.g., PPO-style ratio clipping (DR-GRPO) for the local step-control, convexity properties of the Sinkhorn divergence), not from an intersection of convex geodesic balls.

3.5   The Information Geometry of ENIGMA

The information-geometric view outlined above motivates the combination of training and regularisation methods used in ENIGMA. We couple policy updates (Fisher–Rao), MI-driven semantic alignment on normalised representations (the 'contrastive learning on a hypersphere' lens), and OT-based distributional control (Wasserstein metric mapping across distributions). This geometry-aware composition is especially useful in our CoT-only/MI tie-breaker reward regime: MI with a first-principles constitution supplies the missing process supervision, ensuring that *principled reasoning* augments answer correctness, OT prevents degenerate drifts that CoT-format rewards can otherwise induce.

3.6   Towards universal representations

The Platonic Representation Hypothesis (PRH), noted briefly in our prior work on *ABC Align*, proposes that as models scale, learned representations across architectures and modalities converge toward a shared statistical model of the world, a "platonic" latent structure to which different

encoders increasingly agree [27, 28]. This framing directly motivates the evolution of our work into the "single-objective" view of ENIGMA: if multiple representational pipelines are pulled toward the same latent, then coordinating policy-space motion (Fisher–Rao), semantic binding (InfoNCE/MI), and distributional control (Wasserstein/Sinkhorn) should be feasible on one underlying information manifold.

Recent results add constructive and theoretical support for ENIGMA. Constructively, Jha et al. [29] learn an unsupervised translator that maps text embeddings between unrelated encoders via a universal latent while approximately preserving geometry (cosine/top-1), suggesting invariants that persist under encoder changes. Theoretically, Ziyin and Chuang [33] provide a perfect PRH for embedded deep linear networks: under SGD, two networks of different widths/depths trained on different data converge to identical representations up to rotation. While idealised, this clarifies when "same-up-to-orthogonal-transform" is a reasonable target motivating (though not guaranteeing) the effectiveness of geometry-aware penalties like OT in LLMs.

There are also new empirical results that test PRH in domain-specific settings. Duraphe et al. [34] measure representational convergence across astronomical foundation models (JWST/HSC/Legacy/DESI) using mutual-kNN alignment and report scale-dependent increases in alignment. Adjacent work by Yi, Douady, and Chen [35] provides a theoretical framework for multimodal contrastive learning, showing that under a subspace constraint the modality gap equals the smallest angle between hyperplanes and linking this geometry to pairwise alignment, precisely the kind of structure our InfoNCE-based probes and OT regulariser can monitor. A contemporaneous survey by Lu et al. [36] synthesises cross-modal evidence and emphasizes that objectives and architectures shape how convergence emerges, suggesting that active mechanisms (like our MI binding and OT constraints) may be necessary rather than relying on passive convergence with scale alone. Finally, Gupta et al. [37] and Schnaus et al. [38] demonstrate unpaired multimodal/cross-modal alignment without parallel data, indicating exploitable cross-representational structure whether from PRH-style convergence or shared training distributions. This practical finding motivates ENIGMA's approach of actively coordinating such structure through MI and OT mechanisms.

The growing body of work supporting PRH has implications for ENIGMA. If representations trend toward a shared latent, ENIGMA's composite update can be read as a *product-manifold controller* that accelerates movement toward "platonic directions" consistent with principles: GRPO supplies local Fisher–Rao steps where reward pays off; MI increases kernel-level binding between completions and principles (a PRH-friendly invariant); and entropic OT constrains distributional motion to avoid semantically large but KL-small drifts, exactly the failure modes highlighted by PRH-critical counterexamples [33, 34, 35, 36, 27, 29].

4 Methodology

In this section, we outline our quantitative approach to constitution evaluation, training methods & data, and information geometry probes.

All training and evaluation was performed on single-node GPUs, specifically the NVIDIA A10g with 24GB GDDR6.

4.1 Constitution Evaluation

A central aim of ENIGMA is to connect organisational principles with LLM training and evaluation, by both eliciting *principled reasoning* and constraining model behaviour in a provable way. This section extends our prior work on adapting Constitutional AI for *ABC Align* to include the formal, quantitative evaluation of these principles.

For a candidate principle set C, we estimate three sufficiency signals, each defined in Section C:

1. predictive information ΔNLL (bits/token; token-level perplexity reduction when conditioning on c; Equation (EqC-24))

2. associative information via clean row/column InfoNCE metrics (lower bounds on $I(Y;C \mid X)$ under fixed-K uniform negatives; Equations (EqC-06), (EqC-07), (EqC-09) and (EqC-10))

3. separation (AUC of ΔNLL between positives and negatives, as a measure of discriminative sufficiency; Equation (EqC-25))

We aggregate these into a Sufficiency Index (SI; Equation (EqC-26)) with robust z-scored margins.. Empirically, SI correlates with reduced reward variance, steadier gradient norms, and improved benchmark scores, supporting our use of SI for principle editing prior to model training.

4.1.1 Extension of SAMI-style MI alignment for our training loop

Our MI component borrows the contrastive lens of representation learning: an InfoNCE critic produces a tractable lower bound on MI between pooled representations of the principle and the continuation, $I \geq \log N - L_{\text{InfoNCE}}$. Maximizing this bound increases the association between a principle and responses that satisfy it; comparing the positive/negative margins (or MI bits) flags *leaky negatives* that inadvertently align with desired behaviour. We adopt practical safeguards from the MI literature: the bound saturates at logN and can suffer bias/variance pathologies at high MI. In this way, the evaluator functions as a *SAMI-style* (mutual-information-guided) probe that is compatible with downstream self-supervision or RL.

From a compute requirement and thus training efficiency perspective, our naïve SAMI variant that forms $C \times C$ blocks per question (when multiple principles share a group) increased per-step time by approximately $10\times$ compared with GRPO alone. Our in-batch symmetric row/column InfoNCE restores the step cost to roughly 1.0–$1.2\times$ the GRPO baseline on an NVIDIA A10g (4 completions/prompt), i.e., about an order of magnitude cheaper than the naïve variant. We leave the necessary optimisation of this and impact on training dynamics and downstream task performance as future work.

4.1.2 From constitutional principles to transport-regularised policy updates

In training regimes that combine policy optimisation with distributional regularisation, constitutional principles act as charts that bias updates along geometry-aware directions. Entropically-regularised OT offers a stable, differentiable proxy for Wasserstein geometry, making it natural to penalise unwanted shifts while amplifying principle-consistent ones.

Concretely, starting with the constitutional principles themselves, we find directions (principles) that:

1. increase likelihood of valid chains of thought

2. raise coupling between constraints and outcomes

3. enlarge geometric margins that resist perturbations

In information-geometric terms, robust optimisation formalises adversarial training as worst-case risk over neighbourhoods. This corresponds to controlling motion within metric balls (or entropic OT neighbourhoods) around the data-aligned chart. Thus, constitutions that score high on our 'sufficiency' metrics should simultaneously improve step-by-step reasoning fidelity, normative alignment, and resistance to adversarial prompts, because all three objectives push along the same well-shaped directions of the statistical manifold.

Towards this aim and prior to model training, we evaluate whether a given set of constitutions/principles supplies a sufficiently strong and *selective* learning signal for ENIGMA. To compare principle sets, we calculate the following measures:

- SI (Sufficiency Index). A scalar summary combining (z-scored) MI diag margin, clean MI bound(s), positive-set AUC vs. negative controls, and the median positive ΔNLL (bits/token); higher indicates more principle-encoding evidence in completions.

- ΔNLL (pos). Median per-token NLL improvement on prompts when conditioned on positive principles vs. no/neutral principle.

- AUC (pos vs. neg). Classifier-style separability using the diag-MI statistic as a score.

- MI diag margin. Mean difference between positive and negative diagonal PMI-like scores.

- MI lower bound (bits). Clean row/column metrics, converted to bits where relevant.

- MI-effective (margin mode). A robust aggregate that down-weights outliers and rewards consistent positive/negative separation.

This design deliberately cross-checks multiple views of "principle encoding", avoiding over-reliance on any single bound.

4.1.3 Data and models

We use 1,000 prompts from the CoT-Collection [52] (train split), inserting an editorial standard (positive constitutional principle) and scoring gold rationales/answers. We run two Gemma 3 instruction-tuned base models (1B, 4B) [23], the results of which motivate our selection of the smaller model for efficient ENIGMA experiments.

SAMI explicitly optimizes conditional mutual information between constitutions and model responses, so our MI measures mirror the training signal of ENIGMA's SAMI component. ΔNLL probes whether principles make the right tokens easier to predict and AUC asks whether positive and negative principles are well separated.

4.1.4 Results and Analysis

We compare two principle sets: baseline constitutions vs rewritten constitutions. Both principle sets are then used for our training runs, as 'low SI' and 'high SI' respectively.

The negative principle sets were generated. During the development of our methods, the first negative principle set showed the poor metrics outlined below. We attributed the low effective MI/SI to high lexical overlap and simple negation, e.g. '*yield editorial control'*. For the high effective MI/SI principles, our generation step emphasised 'procedural intent' to improve our metrics. Section B contains both sets & their generation prompts in full, and the systematic generation of negatives towards high effective MI/SI targets remains future work.

Across Gemma-3-1B-IT and Gemma-3-4B-IT, Table 1 summarises our results.

Table 1: Summary of sufficiency signals (higher is better unless noted).

Base model	Principle set	SI ↑	Δ (SI) vs. baseline	Pos. ΔNLL median (bits/tok) ↑	AUC (pos vs. neg) ↑	MI diag margin (pos/neg) ↑	MI lb (bits, pos/neg) ↑	MI-effective (margin mode) ↑
Gemma-3-1B-IT	Baseline	0.715	—	0.123 (≈ **8.2%** perplexity drop)	0.074	6.39 / 3.97	1.40 / 0.62	2.42
	Rewritten	**1.959**	**+1.244**	0.123 (≈ **8.2%**)	**0.272**	**6.39 / −0.045**	1.40 / 0.00	**6.44**
Gemma-3-4B-IT	Baseline	0.582	—	0.0575 (≈ **3.9%** drop)	0.148	6.48 / 4.42	1.43 / 1.07	2.06
	Rewritten	**1.956**	**+1.374**	0.0575 (≈ **3.9%**)	**0.356**	**6.48 / −0.022**	1.43 / 0.00	**6.50**

Notes. Perplexity reduction is $2^{-\Delta\,\text{bits}}$. "AUC" is Mann–Whitney AUC over per-principle ΔNLL; 0.5 denotes no separation; < 0.5 indicates "inverted" separation. Figure 1 visualises these constitution diagnostics for both the low- and high-SI principle sets.

We observe that:

1. Rewriting negatives yields a *large* increase in MI selectivity. The MI diagonal margin for negatives drops from ~4.0–4.4 (baseline 'low SI') to ≈ 0 or negative (rewritten 'high SI'), while positives remain high (~6.4). The InfoNCE MI lower bound for negatives collapses to ~0 bits, producing MI-effective ≈ 6.4–6.5 (vs. ~2.1–2.4 baseline). This is exactly the kind of conditional mutual-information contrast SAMI is designed to exploit. We ablate and verify the impact of this difference in our training setting.

2. ΔNLL for positives is stable but modest (0.06–0.12 bits/tok). This corresponds to an ~4–8% token-level perplexity reduction on gold CoT continuations. This is practically meaningful, but small compared to the MI swing, suggesting that the *association* channel (SAMI) will dominate early training dynamics under ENIGMA.

3. Separation by ΔNLL improves but remains below chance (AUC < 0.5 inverted $\rightarrow$ climbing toward 0.5). After rewriting, AUC rises by ~0.20 absolute for both models (to 0.27–0.36),

(a) Component contributions **(b)** Positive vs. negative bits/AUC **(c)** Paired sufficiency metrics

(d) Effective MI margins **(e)** Sufficiency Index summary

Figure 1: Constitutional sufficiency diagnostics comparing the baseline (low-SI) and rewritten (high-SI) principle sets. Each bar chart shows the relative shift in the metrics that feed the Sufficiency Index, highlighting the MI-driven gains that motivate the rewritten constitution.

but still indicates that many negatives remain "leaky" under the token-level metric. This is consistent with our *editorial-standard* prompt wrapper: even "negative" instructions can sometimes make the gold continuation more likely because the continuation contains the correct answer tokens, as is the case in the KAIST-CoT dataset (we implement a '*leaky-negative* detector' to flag this failure mode).

4. SI gains are driven by MI. With weights wb $=0.6$,wm $=0.3$,ws $=0.1$, the large shift in MI dominates SI: $+1.24$ (1B) and $+1.37$ (4B). The bits component stayed constant; the separation component moved toward 0 but remains negative.

4.2 Training

In our applied setting, we have implemented the optimisation of a single information-geometry objective that balances *reward (GRPO)*, *association (SAMI)*, and *shift (OT)* as a custom trainer ENIGMATrainer and helper scripts in a fork of TRL 0.23.0. In this section, we decompose our single objective into the core components of our methodology.

4.2.1 GRPO core (on-policy RL)

We adopt a standard implementation of GRPO (Group-Relative Policy Optimization), which normalizes returns within groups of completions for the same prompt (conditioned by a positive constitution for all experiments and ablations, per SAMI) and performs clipped policy improvement

akin to PPO, but at the *group* level. We use TRL's GRPOTrainer/GRPOConfig, subclassing it with our ENIGMATrainer, and follow their aggregation, importance sampling and clipping logic.

4.2.2 SAMI auxiliary (sequence-level InfoNCE)

We augment standard GRPO loss with a SAMI auxillary. To define this concretely, let $S_{ij} = \log p_\theta(y_i \mid x_i, c_j)$ denote the *sequence log score* for completion y_i under prompt rendered with principle c_j. We compute two cross-entropies:

- Row InfoNCE: cross-entropy of $\text{softmax}_j(S_{ij})$ with $j = i$.

- Column InfoNCE: cross-entropy of $\text{softmax}_i(S_{ij})$ with label $i = j$.

The SAMI loss is a convex combination with optional per-row/column weights that reflect base reward and gating (entropy quantile). This is a two-sided InfoNCE that encourages (row) each completion to score highest under its own principle-conditioned prompt, and (column) each prompt to score highest for its own completion, tying *principle* and *completion* bidirectionally. InfoNCE provides a variational lower bound on MI; we use it as an *auxiliary* constraint rather than a direct MI estimator in row/column form.

An important limitation to note is that variational MI bounds can be loose and biased in finite samples, we therefore avoid interpreting the auxiliary loss as a calibrated MI estimate and instead use it to shape the manifold and stabilise learning.

4.2.3 "Clean" MI lower-bound metric (ungated, uniform shadows)

To *measure* principle encoding, we log a row-wise InfoNCE-style clean bound per sample where the columns comprise the *true* principle plus K uniformly sampled shadow principles from the positive pool (no gating; uniform negatives). We average over the batch to obtain a simple, stable indicator. A column-symmetric version is also logged. These metrics are calculated while excluding our stabilisation techniques (entropy gating, FR logit preconditioning) and act as diagnostics, not training losses, and empirically correlate with downstream gains in our runs.

4.2.4 Contrastive shaping term

We compute a diagonal PMI-like statistic and add a weak shaping penalty that accentuates high/low ends (quantile mask) while keeping gradients stable (centred target). This term improves separation speed without dominating optimization.

4.2.5 MI-based reward channel (row tie-breaker)

At reward time we add a light, continuous 'tie-breaker' when the binary CoT/XML format rewards saturate and the GRPO learning signal collapses, and we gate by a token-entropy quantile (e.g., 0.8) so only sufficiently decisive rows receive this extra advantage. We track an EMA-based auto-scaler to keep the MI reward's share of total reward near a target.

4.2.6 Entropic Sinkhorn OT (representation regulariser)

We aggregate last-layer hidden states over completion tokens into a per-sequence representation and compute a Sinkhorn divergence between current (adapters on) and reference (adapters off) batches. Entropic OT via the 'geomloss' package is selected for its efficient GPU implementations, bias-reduced "Sinkhorn divergence" behaviour (OT-like at small blur, MMD-like at large blur), and stable gradients. Our implementation uses geomloss' tensorised backend.

4.2.7 ENIGMA

Our unified optimiser follows the manifold direction that satisfies Equation (EqC-17) so extra association is only accepted when it pays off in reward and stays within the OT-bounded shift (with GRPO clipping adding a local trust region). Because the clean InfoNCE bounds in Equations (EqC-09) and (EqC-10) are valid mutual-information lower bounds, observing them out-of-sample provides a lower bound and statistical dependence that completions encode the principles, while the OT term, rather than a KL penalty, governs *where* on the policy manifold those encodings can move.

This single, geometry-aware objective thus unifies reasoning, alignment (make completions carry the principles via MI and row-MI reward) with adversarial robustness (keep shifts bounded in Wasserstein geometry), explaining our empirical pattern of high effective MI with bounded shift even when ΔNLL remains flat under domain mismatch.

4.2.8 Model and data

We briefly cover the base model used and offer our motivation for dataset selection and the 'system prompt' limitation of Gemma 3. See Section A for complete details.

- Base model & adapters. We fine-tune google/gemma-3-1b-it with LoRA.

- Dataset. KAIST CoT-Collection (1.8M CoT rationales across 1,060 tasks) (20k train rows) [52]. Each prompt conditions exactly one positive principle from a YAML constitution; we report results for a low-effective-MI and a high-effective-MI version as the only configuration change

  - We deliberately train on KAIST CoT-Collection on rather than on an editorial/safety corpus. This dataset is domain-mismatched with our editorial principles (math/code vs editorial style), which suppresses trivial lexical overlap between principles and targets. Consequently, changes in gold-answer likelihood (ΔNLL) are expected to be small, while association between principles and responses, quantified by an InfoNCE mutual-information lower bound, can still increase substantially. Observing MI↑ with ΔNLL ≈ 0 is the predicted signature of our information-geometric hypothesis (see Results and Analysis, Table 1: ΔNLL constant at $0.123/0.0575$ bits $\cdot$ tok^{-1} while MI-effective rises from $\sim 2.4 \to 6.4$ and $\sim 2.1 \to 6.5$, respectively): alignment (principle-following) and robustness (bounded distributional shift) are a single optimization on the manifold of policies when combining GRPO (policy objective), SAMI/InfoNCE (association), and Sinkhorn OT (distributional regularisation). This setting follows robustness best practice (evaluate under distribution shift) and provides a reproducible, large-scale probe of the manifold-level effects of out-of-domain constitutional constraints.

- System/user prompts. Gemma 3 does not officially support a 'system prompt' [1], so we supply an instruction specifying our XML/CoT tagging format at the first user turn, which ask the model to respond strictly between <reasoning>...</reasoning><answer>...</answer> tags; the base reward is 1 if the format is exact. We apply the standard Gemma 3 IT chat template.

4.2.9   Trainer configuration

See Section A for complete implementation details and additional hyperparameters. These values are specific to the training runs reported here.

4.2.10   ENIGMA Training

We set the following parameters for both our ENIGMA runs across both sets of constitutions, 'high SI' and 'low SI'.

- Generation. 4 sampled completions per prompt (temperature=1.0, top_p=0.95, top_k=64, repetition penalty 1.1).

- RL algorithm. GRPO with dr_grpo loss (sequence-level ratio clipping, group-wise advantage centering/scaling), no KL to reference (beta=0). We use sequence-level importance weights, mask_truncated_completions=True, and epsilon=0.1.

- SAMI auxiliary. In-batch symmetric InfoNCE between prompts (question+principle) and completions with row/column ratio annealed from 0.7/0.3 to 0.5/0.5 over the first 10% of steps; warmup ramps mi_weight=0.05 from 0 to full over 50 steps. InfoNCE scores use length or logit-Fisher normalisation.

- Row-MI reward channel. A gated dense reward converts the row log-softmax at the positive principle (with K=2 shadow principles) to [0,1] via a sigmoid (slope 2.5), weighted by 0.15. The reward is:

  - Entropy-gated (keep rows below 80th percent sequence-entropy), and
  - Format-gated (after approximately 30% of MI warmup, only completions passing the XML format).
  - An EMA autoscaler keeps MI reward near 20% of total reward magnitude on average [60].

- Sinkhorn OT regulariser. After a 200-step warmup, we add $\lambda_{OT}S_\varepsilon$ between normalised completion-token hidden-state means of the current policy and the adapter-disabled reference (ot_weight=0.01, blur=0.12, scaling=0.8).

4.2.11   GRPO-only Training

We perform ablations with two variants of GRPO-only runs. Neither MI nor OT is used.

For GRPO CoT, we use the same XML-format binary reward function we use for both ENIGMA runs. For GRPO CoT+, we use the XML format reward with a Gaussian noise 'tie-breaker' reward, as a stand-in for our MI equivalent in ENIGMA.

All other hyperparameters are identical.

4.3 Information Geometry probes

See Section A for additional implementation details.

We log two families of probes to connect training dynamics to our single-objective information-geometric hypothesis:

1. Output distribution proximity (last token).
   Bhattacharyya angle, Hellinger distance, and Jensen–Shannon divergence between current and reference logits (Equations (EqC-18) to (EqC-20)).

2. Capacity/flatness proxies.
   Fréchet distance between batches of hidden summaries and effective dimensionality (effective rank; participation ratio) to monitor over-/under-concentration in representation space (Equations (EqC-21) to (EqC-23)).

These probes are not optimised directly (aside from the OT term) and serve as training-time correlates that we observe to move with our MI/sufficiency signals.

5 Model Evaluation

5.1 Benchmarks

We use lm-evaluation-harness (v0.4.9) [44] with fixed seeds/recommended Gemma 3 IT decoding settings and a vLLM (10.1.2) [53] back-end for efficient completions. For benchmarks, we use GPQA (main) [30] and TruthfulQA [31], Both benchmarks are the 'generate_until' variants with the same prompt used for training that contains CoT tags, <reasoning>.

Results are from complete ENIGMA training with low/high effective MI constitution samples for LoRA checkpoint 2000 (merged back into base model)

5.2 Results

5.3 Benchmark performance vs. Gemma-3-1B-IT baseline

Table 2: Benchmark performance relative to the Gemma-3-1B-IT base model.

Run	GPQA flex-EM	Δ vs. base	TruthfulQA BLEU-acc	Δ vs. base
Baseline	0.1362	—	0.3807	—

(a) Absolute GPQA performance

(b) GPQA gains over the base model

(c) Absolute TruthfulQA BLEU-accuracy

(d) TruthfulQA gains over the base model

Figure 2: Benchmark performance across training variants. Each bar chart reports the absolute score or improvement relative to the Gemma-3-1B-IT baseline for GPQA (top row) and TruthfulQA (bottom row). The plots highlight the complementary behaviour of ENIGMA High-SI (joint gains) and Low-SI (GPQA-only gains).

Table 2: Benchmark performance relative to the Gemma-3-1B-IT base model.

Run	GPQA flex-EM	Δ vs. base	TruthfulQA BLEU-acc	Δ vs. base
GRPO CoT	0.1406	+0.0045	0.3611	-0.0196
GRPO CoT+	0.1830	+0.0469	0.3390	-0.0416
ENIGMA High SI	**0.2054**	**+0.0692**	**0.5018**	**+0.1212**
ENIGMA Low SI	**0.2366**	**+0.1004**	**0.2399**	**-0.1408**

Figure 2 summarises the GPQA and TruthfulQA benchmark outcomes that accompany the aggregate scores in Table 2.

5.3.1 Notes & Interpretation

- **ENIGMA High SI** significantly improves both tasks; **ENIGMA Low MI** significantly harms TruthfulQA while improving GPQA.

5.4 Training dynamics at step 2000

We report $I = \log(K+1) - L$. In the large-sample limit with unbiased negatives this quantity is ≥ 0. In practice, (a) using a small K, (b) gating to the "clean" subset, and (c) score mis-calibration can introduce bias, so I can be slightly negative early in training or in ablations. We keep the sign to make "worse-than-chance" association visible; interpret magnitudes *relatively* across runs/steps rather than as an absolute MI estimate.

Table 3: Training diagnostics at step 2000 for each variant.

Run	reward_std	grad_norm	entropy	MI row bound (clean, K = 2, nats)	MI col bound (clean, K = 2, nats)	MI row-col gap	OT
GRPO CoT	**0.00**	**0.00**	0.206	-0.0188	-0.1327	+0.114	0.000
GRPO CoT+	0.50	2.282	0.351	+0.0078	-0.0085	+0.016	0.000
ENIGMA High SI	0.021	1.430	**0.112**	**+0.0941**	-0.0461	**+0.140**	0.036
ENIGMA Low SI	0.029	1.185	0.139	**+0.1018**	~0.000	+0.102	**0.064**

5.4.1 Notes & Interpretation:

MI row/col bound (clean). We treat principle identification as a small retrieval game. *Row* asks: given a completion, can we tell which principle it reflects better than two shadow principles? *Column* asks the inverse: given a principle, can we pick out its completion among distractors? We compute a standard InfoNCE lower bound on $I(Y; C \mid X)$ in each direction and report it on the clean subset only (low entropy, valid XML), which removes degenerate strings and isolates *principled reasoning.* With K=2 negatives, bounds live in $[-\infty, \log 3]$ nats; values near 0 indicate chance-level association, positive values indicate that generated reasoning preferentially encodes the stated principles. We also report the row–col gap, which is large when completions show surface compliance (row ↑) without uniquely binding to the principle (col ≈ 0).

- **Ablations.** *GRPO CoT* shows no learning signal and slightly negative row-MI:format compliance without principle content. *GRPO CoT+* restores a weak signal (row ≈ 0.008 nats), but transfer is limited, matching its small IG shifts.

- **ENIGMA High-SI.** Row-MI clean ≈ 0.094 nats (≈ 0.136 bits) with a modest JS/Hellinger shift indicates structured movement toward principle-consistent distributions rather than broad stylistic drift. Gains on TruthfulQA are consistent with *column* MI remaining slightly negative; principles guide completions but are not yet uniquely identifying them among distractors.

- **ENIGMA Low-SI.** Row-MI clean ≈ 0.102 nats with minimal output shift and a larger OT cost suggests the model learns *surface* correlations that help GPQA but hurt TruthfulQA; "style over substance." The larger row–col gap flags this pattern.

- We observe that larger row–col gaps correlate with failures on adversarially phrased TruthfulQA items, suggesting that column-direction identifiability is a better proxy for robustness than row alone.

5.4.2 Information-geometry probes

Table 4: Information-geometry probe summary at step 2000.

Run	Bhattacharyya angle	Hellinger	JS
GRPO CoT	**0.742**	**0.509**	**0.234**
GRPO CoT+	0.362	0.251	0.103
ENIGMA High SI	0.371	0.245	0.149
ENIGMA Low SI	0.099	0.070	0.008

5.4.3 Notes & Interpretation

- The *format-only* ablation induces the largest last-token shift (over-fitting to CoT tag/style).

- **ENIGMA High SI** shows a moderate but structured shift; *lowmi* shows a minimal shift, consistent with better GPQA (domain general reasoning) but worse TruthfulQA (truthfulness/robustness).

5.5 Constitution Sufficiency Index (SI)

Now, we can connect our constitution evaluation metric, SI, with benchmark results and training dynamics.

- **Low-SI constitutions:** SI ≈ 0.715 (components: bits $= 0.123$, MI ≈ 2.42, separation $= -0.85$)

- **High-SI constitutions: SI** ≈ 1.959 (MI ≈ 6.44, better separation), with the same bits term.

- Benchmarks results are consistent with these measurements: **high-SI** $\rightarrow$ truthfulness gains; **low-SI** $\rightarrow$ truthfulness loss despite GPQA gains. This is an empirical validation that SI is a useful metric for the selection and curation of principle sets.

5.6 Discussion

Recalling our core hypothesis, that reasoning, alignment, and adversarial robustness are a *single* optimisation objective; SAMI can shape rewards (representation space) and constrain parameters (auxiliary loss), while OT regularises transport between distributions.

Our results provide the following evidence to support our claim:

1. Representation MI $\uparrow \rightarrow$ Truthfulness $\uparrow$ (and robustness)
   Across steps, **ENIGMA High MI** pushes the row-wise MI lower bound from near-zero

early to positive at step 2000 (and higher later), while ENIGMA Low SI is also positive but paired with larger OT core and worse TruthfulQA. Only **ENIGMA High MI** shows *joint* improvements, consistent with the idea that what MI binds to (a high-SI constitution) matters more than "MI magnitude" alone.

2. Optimisation needs stochasticity + relative advantages, but the *content* signal must be MI-grounded.
   The Gaussian 'jitter' added to GRPO CoT+ converts a degenerate reward to an informative signal, as GRPO relies on variance in group-relative advantages. Yet without MI shaping the learned policy pursues style/format rather than truthfulness.

3. OT regularisation co-controls *what* moves and *how much*.
   Entropic Sinkhorn distances give stable transport with a temperature/ε that limits overconfident shifts. Our Enigma Low SI run has a larger OT core at step 2000, consistent with "moving mass" to easy-to-satisfy surface features (format/compliance) rather than semantic truth.

Together, these observations support our manifold view: SAMI (InfoNCE) = geometry-aware content binding, OT = geometry-aware movement, GRPO = variance-driven policy improvement; the *high-SI constitution* provides the signal that aligns all three.

5.7 Benchmark selection and qualitative evaluation

Our aim here is to validate our core hypothesis empirically, and connect organisational standards and quantifiable measures and offer a practical implementation guide that addresses gaps in the development community. It is not to demonstrate the optimal performance characteristics of our methods against benchmarks. We use benchmarks primarily to validate the relationship between our constitution evaluation metrics, training metrics, and downstream task performance.

However, beyond the results reported above, a qualitative view of model outputs offers insight into how these performance characteristics manifest in a clear and simple way, despite their limited training duration (2000 steps) and minimal compute.

Taking an example from the dataset:

TruthfulQA prompt.

> **Question:** What happens to you if you eat watermelon seeds?

Scoring references. TruthfulQA accepts the following completions as correct:

- Nothing happens.
- You eat watermelon seeds.
- The watermelon seeds pass through your digestive system.
- You will not digest the watermelon seeds.
- The watermelon seeds will be excreted.
- I have no comment.

Model completions. We compare the baseline Gemma-3 model to ENIGMA variants with and without reasoning traces:

> *Gemma-3 baseline (CoT):* You might experience mild discomfort, such as a slight tingling sensation.

> *ENIGMA Low MI (CoT):* It doesn't happen.

> *ENIGMA Low MI (no CoT):* It happens to you that you get a tummy ache.

> *ENIGMA High MI (CoT):* I have no comment.

> *ENIGMA High MI (no CoT):* It's generally harmless to eat watermelon seeds.

Gemma3, even with our CoT prompt, provides an incorrect answer. ENIGMA High MI's completion is a refusal, which scores correctly on TruthfulQA, without a CoT it still provides correct answer. Enigma Low MI's completion is a failure of alignment that reads as an attempted denial.

The limitations of current approaches to LLM evaluation are discussed in the literature. We have selected TruthfulQA both to build on our prior work, and as a measure of both alignment and adversarial robustness because by design, it encourages models to make 'imitative falsehoods' across a broad range of domains. It balances an evaluation of both alignment and adversarial robustness and we posit that it tests something more fundamental about reasoning and language, which we quantify through the lens of information geometry via both our SI metric and use of OT.

5.8 Information Geometry Analysis of Trained Adapters

We now inspect the properties of the trained adapters across all runs, to further quantify the relationship between our constitution evaluation metric SI, training dynamics, and task performance, from an information geometry perspective. This analysis was conducted using 300 random samples of both GPQA (main) TruthfulQA, and measuring manifold changes, across checkpoints 500, 1000, and 2000.

For each dataset, across each training run, we plot various measures across three panels.

5.8.1 Panel A: Training-path triptych

Top: Δ(diagonal SAMI MI) vs. checkpoint ($0 \rightarrow 500 \rightarrow 1000 \rightarrow 2000$). Because our SAMI objective is an InfoNCE-style bound, *less-negative or increasing* values indicate a stronger certified association ("binding") between principle text and gold continuations; decreasing values indicate weakening association.

Middle: Cumulative Fisher–Rao (FR) path length L (solid) vs. endpoint FR geodesic dgeo (dashed). When L>dgeo , training takes a *non-geodesic* route on the statistical manifold; the ratio L/dgeo summarizes FR-efficiency.

Bottom: Discrete turning angles between successive segments approximate curvature: larger angles $\rightarrow$ more tortuous paths. FR is the intrinsic Riemannian metric for distributions (natural gradient), and on the simplex its geodesic distance is the Bhattacharyya/statistical angle after the p map, so these path quantities are coordinate-free.

5.8.2   Panel B: Landscape triptych

Left: SAMI diagonal MI heatmap (IC-aligned) over (α, β)-perturbations of the decoding distribution. Middle: FR distance heatmap under teacher forcing.

Right: MI heatmap with FR iso-contours overlaid. Where MI ridges run with low/flat FR contours, the training can increase MI without moving far on the manifold (FR-efficient); where ridges cut across steep contours, MI gains are FR-costly. (FR background; InfoNCE bound.)

5.8.3   Panel C: ICMI panel

Left: Mean aligned L-matrix (row-wise InfoNCE) across principle-completion pairs; a bright diagonal indicates that principles reliably bind to their own gold continuations; bright off-diagonals signal cross-binding.

Right: Histogram of diagonal MI per row quantifies the spread: a right-shift (less negative) distribution indicates stronger binding across many rows; heavy left tails indicate rows where association is weak or inverted. (InfoNCE bound.)

5.9   GPQA – ENIGMA High/Low SI, GRPO CoT, GRPO CoT+

GPQA rewards hard reasoning and domain synthesis; GRPO CoT+ can raise MI here, but by pushing along stiff directions; ENIGMA-High-SI remains curved but with more principled (ICMI-certified) binding.

5.9.1   ENIGMA – High SI

Panel A of Figure 3 shows $L/d_{\mathrm{geo}} \approx 2.76$, indicating a curved, non-geodesic trajectory with MI dipping negative by 2k steps (with only a brief recovery at 1k). Panel B highlights moderate Fisher–Rao anisotropy where MI ridges partially misalign with low-cost contours, so sizeable MI gains remain possible but incur additional transport. Panel C displays a clear diagonal with a right-shifted histogram, confirming robust though non-uniform binding across GPQA rows.

5.9.2   ENIGMA – Low SI

Panel A of Figure 4 tracks $L/d_{\mathrm{geo}} \approx 1.72$, closer to a geodesic but with MI still declining by step 2000. Panel B reveals anisotropy below one, meaning MI ridges align with flatter Fisher–Rao directions and offer more efficient moves than High-SI. Panel C shows a weaker diagonal and heavier left tail than the High-SI run, signalling incomplete principle binding across GPQA rows.

5.9.3   GRPO CoT

Panel A of Figure 5 shows $L/d_{\mathrm{geo}} \approx 1.85$, almost geodesic yet still trending toward lower MI by step 2000. Panel B exhibits low anisotropy with MI ridges largely aligned to shallow Fisher–Rao contours, implying limited structured guidance. Panel C illustrates a diffuse diagonal and a histogram concentrated near negative MI, highlighting weak per-row binding under the format-only reward.

5.9.4 GRPO CoT+

Panel A of Figure 6 records $L/d_{\text{geo}} \approx 2.81$ with strong anisotropy ($\alpha/\beta \approx 5.24$), so MI gains demand large Fisher–Rao moves. Panel B shows MI ridges running across steep contours, signalling FR-costly improvements. Panel C sharpens the diagonal relative to GRPO CoT but retains broad spread, indicating uneven gains concentrated in a subset of prompts.

5.10 TruthfulQA – ENIGMA High/Low SI, GRPO CoT, GRPO CoT+

TruthfulQA penalises spurious associations that mimic common falsehoods. Only ENIGMA-High-SI shows consistent, row-wise binding gains without paying large FR costs, matching its large truthfulness lift; GRPO-only either moves too little (CoT) or in the wrong directions (CoT+.

5.10.1 ENIGMA – High SI

Panel A of Figure 7 ends slightly positive by step 2000 after a dip at 500 and recovery at 1000, even though the Fisher–Rao path remains long and curved. Panel B highlights moderate anisotropy with MI ridges occupying relatively flat contours, enabling FR-efficient MI improvements. Panel C presents a strong diagonal and right-shifted histogram, aligning with the TruthfulQA gains reported for this run.

5.10.2 ENIGMA – Low SI

Panel A of Figure 8 drops in MI by step 2000 despite similar curvature to the High-SI run. Panel B reveals MI ridges that frequently cross steep Fisher–Rao contours, making MI improvements costly. Panel C depicts a weaker diagonal with heavier left tail, signalling insufficient binding and matching the degraded TruthfulQA results.

5.10.3 GRPO CoT

Panel A of Figure 9 yields $L/d_{\text{geo}} \approx 1.67$ with a slight MI increase, reflecting modest yet FR-cheap progress. Panel B indicates low anisotropy and mild alignment between MI ridges and flat contours. Panel C retains a diffuse diagonal with less right shift than ENIGMA High-SI, showing weaker overall binding even when MI moves in the right direction.

5.10.4 GRPO CoT+

Panel A of Figure 10 yields $L/d_{\text{geo}} \approx 2.52$ with MI decreasing, revealing that stochastic jitter pushes the policy along inefficient Fisher–Rao directions. Panel B confirms MI ridges frequently cross steep contours, so improvements are FR-costly. Panel C shows a weaker diagonal and broader negative spread than ENIGMA High-SI, reflecting degraded binding.

5.11 Discussion

Across datasets and runs, the three panels tell a coherent story: ENIGMA with a sufficient principle set (High-SI) reshapes the FR landscape so that directions which increase the InfoNCE MI bound (Panel B) are accessible along relatively flat FR contours, enabling positive or preserved MI at step 2k even on curved paths (Panel A), and producing a clear ICMI diagonal (Panel C). In contrast, GRPO-only either makes small MI gains with weak binding (CoT) or pursues FR-costly, anisotropic directions (CoT+) that help GPQA but hurt truthfulness.

These IG measurements are consistent with our overarching hypothesis: reasoning, alignment, and robustness can be treated as one optimszation problem on the information manifold. When the principle set is sufficient and enforced via an InfoNCE-style auxiliary (SAMI), ENIGMA's unified optimisation objective aligns reward-preferred moves with FR-efficient principle-binding moves.

6 Limitations

We have limited our experiments to a single model, Gemma 3 1B, dataset (KAIST-CoT), and set of human-written principles (with two sets of generated negatives). Our aim in this work is to introduce a novel approach to simultaneously improving LLM reasoning, alignment and robustness in a fundamental way, that is ultimately directed by human-written guidelines or principles for LLM behaviour. We have not exhaustively evaluated performance of our methods on the many open models & datasets. While significant effort has been spent on developing and adapting new and complimentary techniques and metrics, we leave further optimisation to those who choose to apply ENIGMA in their research or domain-specific application.

7 Future Work

We believe ENIGMA will scale to much larger models, though have not run these experiments ourselves. Our MI-based reward tiebreaker is designed to sustain a learning signal as format rewards saturate and expect this will facilitate scaling our technique to 100B+ parameters. Additional work will be needed to stabilise training on MoE architectures. There is growing literature on the use of synthetic data for pretraining small LLMs that demonstrate strong reasoning performance [32], and have covered the topic of synthetic data specifically in our prior work on *ABC Align*.

Regarding our information geometric objective specifically, Wasserstein/Sinkhorn trust-regions support our general approach in ways we have outlined, however token-level costs and their relationship to true distributional semantics remains future work. The many tools and techniques from Information Geometry we have leveraged and the growing evidence to support the PRH point to a research future of convergence, where the inductive biases of geometry and fundamental interventions taking an information geometric perspective enable may improve LLM performance in a general way. We also leave such explorations to others, and as future work for ourselves, in alignment with organisational objectives.

8 Conclusion

Our evidence suggests that key behavioural characteristics of LLMs, reasoning, alignment, robustness, can be jointly optimised by a single information-geometric objective that shapes the underlying information manifold in a measurable way. The emerging evidence for PRH, from unsupervised encoder translation to formal convergence results offer additional support for our single objective. The manifold shaping improves LLM performance in a manner desirable given the context of their deployment; GRPO improves the policy where it pays off, SAMI couples reasoning to principles, and Sinkhorn OT limits harmful geometry shifts. 'Clean' InfoNCE metrics captured during training make this coupling measurable and provide a lower-bound that confirms ENIGMA models encode constitutional principles in their CoT in a manner that is falsifiable. The SI metric we produce as part of our 'constitution evaluation' enable organisations to collaborate and define constitutions before any LLM training, providing confidence that high-SI principle sets reliably predict better training dynamics and downstream task performance. This connects human-written standards directly with model training inputs, model structure, and model behaviour.

We view trusted capability, rather than pure capability alone, as a critical component of future LLM deployments. ENIGMA provides a way to quantify and uphold this trust throughout the LLM lifecycle.

9 References

References

[1] Google AI. (2025). *Gemma formatting and system instructions.* Documentation. https://ai.google.dev/gemma/docs/formatting

[2] Australian Broadcasting Corporation (ABC). (2025). *Editorial Policies.* https://about.abc.net.au/publication/editorial-policies/

[3] Shao, Z., Yu, H., Gong, Y., Zhang, S., & Zhou, J. (2024). DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv:2402.03300.*

[4] Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust Region Policy Optimization. *arXiv:1502.05477.*

[5] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. *arXiv:1707.06347.*

[6] Li, W., & Montúfar, G. (2018/2021). Natural Gradient via Optimal Transport. *arXiv:1803.07033.*

[7] Arbel, M., Gretton, A., Li, W., & Montúfar, G. (2019/2020). Kernelized Wasserstein Natural Gradient. *arXiv:1910.09652.*

[8] Fränken, J.-P., Qu, X., Grewe, D., Schölkopf, B., & Buettner, F. (2024). Self-Supervised Alignment with Mutual Information: Learning to Follow Principles without Preference Labels. *arXiv:2404.14313.*

[9] Bai, Y., Kadavath, S., Kundu, S., Askell, A., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073*.

[10] Ambrosio, L., Brué, E., & Semola, D. (2021). Lecture 18: An Introduction to Otto's Calculus. In *Lectures on Optimal Transport*. Springer.

[11] Peyré, G., & Cuturi, M. (2019). *Computational Optimal Transport: With Applications to Data Science. Foundations and Trends in Machine Learning*, 11(5–6), 355–607.

[12] Melnyk, I., Mroueh, Y., Belgodere, B., Rigotti, M., Nitsure, A., Yurochkin, M., Greenewald, K., Navratil, J., & Ross, J. (2024). Distributional Preference Alignment of LLMs via Optimal Transport. *NeurIPS 2024 (OpenReview)*. arXiv:2406.05882.

[13] Chen, X., Zhu, W., Qiu, P., Wang, H., Li, H., Wu, H., Sotiras, A., Wang, Y., & Razi, A. (2025). Prompt-OT: An Optimal Transport Regularization Paradigm for Knowledge Preservation in Vision-Language Model Adaptation. *arXiv:2503.08906*.

[14] Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-I., Trouvé, A., & Peyré, G. (2019). Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. *AISTATS 2019 (PMLR 89)*, 2681–2690.

[15] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022/2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv:2201.11903*.

[16] Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2023). Let's Verify Step by Step. *arXiv:2305.20050*.

[17] Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S., Leike, J., Kaplan, J., & Perez, E. (2025). *Reasoning Models Don't Always Say What They Think*. Anthropic technical report.

[18] Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., & Tian, Y. (2024). Training Large Language Models to Reason in a Continuous Latent Space. *arXiv:2412.06769*.

[19] Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A., et al. (2025). Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety. *arXiv:2507.11473*.

[20] Poole, B., Ozair, S., van den Oord, A., Alemi, A. A., & Tucker, G. (2019). On Variational Bounds of Mutual Information. *arXiv:1905.06922*.

[21] Martens, J. (2020). New Insights and Perspectives on the Natural Gradient Method. *Journal of Machine Learning Research*, 21, 1–76.

[22] Davidson, T. R., Falorsi, L., De Cao, N., Kipf, T., & Tomczak, J. M. (2018). Hyperspherical Variational Auto-Encoders. *arXiv:1804.00891*.

[23] van den Oord, A., Li, Y., & Vinyals, O. (2018/2019). Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748*.

[24] Nielsen, F. (2022). An Elementary Introduction to Information Geometry (updated version). *Entropy*, preprint/PDF v2.

[25] Peyré, G., & Cuturi, M. (2018/2020). Computational Optimal Transport. *arXiv:1803.00567.*

[26] Ambrosio, L., Gigli, N., & Savaré, G. (2008). *Gradient Flows: In Metric Spaces and in the Space of Probability Measures.* Birkhäuser.

[27] Huh, M., Li, A., Isola, P., & Geiger, A. (2024). The Platonic Representation Hypothesis. *arXiv:2405.07987.*

[28] Huh, M., Li, A., Isola, P., & Geiger, A. (2024). Position: The Platonic Representation Hypothesis. *Proceedings of Machine Learning Research*, Vol. 235.

[29] Jha, R., Zhang, C., Shmatikov, V., & Morris, J. X. (2025). Harnessing the Universal Geometry of Embeddings. *arXiv:2505.12540.*

[30] Rein, D., Li Hou, B., Cooper Stickland, A., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R. (2023). GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv:2311.12022.*

[31] Lin, S., Hilton, J., & Evans, O. (2021). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv:2109.07958.*

[32] Zhao, C., Chang, E., Liu, Z., Chang, C.-J., Wen, W., Lai, C., Cao, S., Tian, Y., Krishnamoorthi, R., Shi, Y., & Chandra, V. (2025). MobileLLM-R1: Exploring the Limits of Sub-Billion Language Model Reasoners with Open Training Recipes. *arXiv:2509.24945.*

[33] Ziyin, L., & Chuang, I. (2025). Proof of a perfect platonic representation hypothesis. arXiv:2507.01098.

[34] Duraphe, K., Smith, M. J., Sourav, S., & Wu, J. F. (2025). The Platonic Universe: Do Foundation Models See the Same Sky? arXiv:2509.19453.

[35] Yi, L., Douady, R., & Chen, C. (2025). Decipher the Modality Gap in Multimodal Contrastive Learning: From Convergent Representations to Pairwise Alignment. arXiv:2510.03268.

[36] Lu, J., Wang, H., Xu, Y., Wang, Y., Yang, K., & Fu, Y. (2025). Representation Potentials of Foundation Models for Multimodal Alignment: A Survey. arXiv:2510.05184.

[37] Gupta, S., Sundaram, S., Wang, C., Jegelka, S., & Isola, P. (2025). Better Together: Leveraging Unpaired Multimodal Data for Stronger Unimodal Models. arXiv:2510.08492.

[38] Schnaus, D., Araslanov, N., & Cremers, D. (2025). It's a (Blind) Match! Towards Vision–Language Correspondence without Parallel Data. In *CVPR 2025* (proc.). arXiv:2503.24129.

[39] Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., & Awadallah, A. (2023). Orca: Progressive Learning from Complex Explanation Traces of GPT-4. arXiv:2306.02707.

[40] Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., ... Levy, O. (2023). LIMA: Less Is More for Alignment. arXiv:2305.11206.

[41] Hong, J., Lee, N., & Thorne, J. (2024). ORPO: Monolithic Preference Optimization without Reference Model. arXiv:2403.07691.

[42] Seneque, G., Ho, L-H., Kuperman, A., Erfanian Saeedi, N., & Molendijk, J. (2024). ABC Align: Large Language Model Alignment for Safety & Accuracy. arXiv:2408.00307.

[43] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Finn, C., & Khodak, M. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290.

[44] Biderman, S., Schoelkopf, H., Sutawika, L., Gao, L., . . . Zou, A. (2024). Lessons from the Trenches on Reproducible Evaluation of Language Models. arXiv:2405.14782.

[45] Bhattacharyya, A. (1943). On a Measure of Divergence between Two Statistical Populations. *Bulletin of the Calcutta Mathematical Society*, 35, 99–109.

[46] Lin, J. (1991). Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1), 145–151.

[47] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *NeurIPS 2017* (introduces Fréchet Inception Distance).

[48] Roy, O., & Vetterli, M. (2007). The Effective Rank: A Measure of Effective Dimensionality. *EUSIPCO 2007*.

[49] Recanatesi, S., et al. (2022). Dimensionality in Recurrent Neural Networks: A Scale-dependent Measure. *Nature Communications*, 13, 7363. (on participation ratio / effective dimensionality).

[50] Amari, S.-I., & Nagaoka, H. (2000). Methods of Information Geometry. AMS/Oxford.

[51] Miyamoto, H. K., et al. (2023). On Closed-Form Expressions for the Fisher–Rao Distance. arXiv:2304.14885. (contains the local KL $\approx \frac{1}{2}d_{\mathrm{FR}}^2$ relation and the Bhattacharyya/statistical angle formula).

[52] Lyu, C. J., Park, H. S., Park, E., Ban, T.-W., & Seo, M. (2024). CoT-Collection: A Large Language Model Dataset with 1.8M Chain-of-Thoughts. arXiv:2412.06851.

[53] Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., . . . Stoica, I. (2023). Efficient Memory Management for LLM Serving with PagedAttention (vLLM). arXiv:2309.06180.

[54] Hugging Face. (2025). TRL – GRPO Trainer (documentation). https://huggingface.co/docs/trl/main/en/grpo_trainer (accessed 2025-10-12).

[55] Ramesh, B., Meng, K., Zhu, A., Castricato, L., et al. (2024). Group Robust Preference Optimization: Improving Robustness with Long-Tail Preferences. arXiv:2410.01166.

[56] National Security Agency Artificial Intelligence Security Center (AISC). (2025). NSA's AISC Releases Joint Guidance on the Risks and Best Practices in AI Data Security. Press release. https://www.nsa.gov/Press-Room/Press-Releases-Statements/

[57] National Artificial Intelligence Centre, Department of Industry, Science and Resources (Australia). (2024). Voluntary AI Safety Standard. Government publication. https://www.industry.gov.au/publications/voluntary-ai-safety-standard

[58] AI Security Institute (AISI), Department for Science, Innovation and Technology (UK). (2025). The AI Security Institute. https://www.gov.uk/government/organisations/ai-safety-institute

[59] Chen, Z., Badrinarayanan, V., Lee, C.-Y., & Rabinovich, A. (2017). GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks. arXiv:1711.02257

[60] Wang, S., Yu, L., Gao, C., Zheng, C., Liu, S., et al. (2025). Beyond the 80/20 Rule: High-Entropy Minority Tokens Drive Effective Reinforcement Learning for LLM Reasoning. arXiv:2506.01939

A   Implementation Details

A.1   Overview

ENIGMA augments TRL's GRPO loop with two information–theoretic additions: SAMI (self-supervised alignment via mutual information) and an entropic Sinkhorn optimal transport regulariser, so that *reasoning*, *alignment*, and *robustness* are optimised on the same information manifold. Concretely:

1. **Online RL.** We use TRL 0.23.0's GRPOTrainer [54] (loss type dr_grpo) which performs group-relative policy optimization on K completions per prompt with advantage normalization and PPO-style clipping. Rewards are centered per group and, unless scale_rewards="none", scaled by the group's standard deviation; ratio clipping is applied at the sequence level (importance sampling level "sequence").

2. **SAMI: two roles.**

   a. **Row/column InfoNCE auxiliary** on (prompt with constitution) ↔ (completion) pairs encourages *principled* reasoning traces to be encoded in the policy by maximizing a symmetric mutual-information lower bound (row- and column-wise). Implementation uses the standard InfoNCE bound $I \geq \log M + \log p_+ - \log \sum_j p_j$, following CPC/InfoNCE and its analysis.

   b. **Row-wise MI reward channel** adds a completion-dependent, *format-gated* dense reward derived from the same row InfoNCE ingredients, to shape GRPO's credit assignment toward completions that best encode the active principle.

3. **Entropic Sinkhorn OT**

   a. We use the Sinkhorn divergence S_ε between empirical measures of L2-normalised sequence-level hidden summaries (current policy vs. adapters-off reference). S_ε is a symmetric, positive-definite, smooth loss that is convex in each argument and metrizes convergence in law; as $\varepsilon \to 0$ it approaches W_2^2, and as $\varepsilon \to \infty$ it converges to a kernel MMD^2 induced by $k_\varepsilon(x, y) = \exp\left(-\|x - y\|^2/\varepsilon\right)$, up to convention-dependent constants in the RKHS norm. We implement S_ε via GeomLoss' `SamplesLoss` (tensorised backend).

4.

We also log a set of **information-geometry probes** (Bhattacharyya angle, Hellinger, JS divergence on last-token distributions; Fréchet distance between hidden-state Gaussians; effective rank and participation ratio; Equations (EqC-18) to (EqC-23)) to monitor training geometry and alignment stability.

A.2 Data, prompt rendering, and strict format reward

Dataset. We use *KAIST CoT-Collection* (kaist-ai/CoT-Collection, split=train, take=20k). Each example supplies source (question) and target (gold rationale/answer).

Constitution conditioning. For each training item, one **positive** principle is sampled from a YAML pool and injected into the user message as:

Constitution:
{principle}

{tag_preamble}

Question:
{source}

The chat template comes from the model tokeniser; we enforce left padding and a strict EOS/EOT at "<end_of_turn>" for Gemma-3 IT. Prompts exceeding max_prompt_length=256 are filtered. Only one positive principle is conditioned per row (no multi-principle grouping on a single question in the mapped dataset).

Format-only base reward. The scalar reward function xml_format_reward returns 1.0 iff the completion is exactly:

<reasoning>... </reasoning><answer>... </answer>

(anchored; exactly one pair of tags; no extra leading/trailing content). Otherwise 0.0. This makes the base task *hard-format* CoT and leaves content supervision to MI/OT.

Note. This reward deliberately does not inspect the content of <reasoning> or <answer>; content shaping is provided by MI components (below).

A.3 GRPO integration and loss

We keep TRL's GRPO generation/scoring pipeline with sequence-level importance weighting and use the Dr.-GRPO loss type (loss_type="dr_grpo"), which divides by a fixed constant (max completion length) to remove length bias; KL-to-reference is off by default (beta=0). Ratio clipping is applied at the sequence level ($\varepsilon = 0.1$). See TRL GRPO docs for the precise forms and options.

Advantages. TRL computes per-sample rewards, centers them by the group mean, and scales by the group std when scale_rewards="group" (our setting), improving stability under group sampling.

Old log-probabilities. We cache old_per_token_logps at generation time to make importance sampling correction meaningful if generation and optimization are misaligned; otherwise we fallback to per_token_logps.detach() per TRL's implementation.

A.4 MI probes & metrics

A.4.1 Sequence score critic used everywhere (definitions used by meters and auxiliary)

We score a completion y under a rendered prompt (x, c) using the length-normalised sequence log-likelihood $s(y \mid x, c) = (1/|y|) \sum_t \log p_\theta(y_t \mid y_{<t}, x, c)$. Unless stated otherwise, we do not apply Fisher/logit weighting in the diagnostics; length normalisation alone was stable in our runs. (The auxiliary may optionally use a Fisher-weighted variant; see below.)

A.4.2 Row/column InfoNCE diagnostics (naming and bounds)

We report two contrastive identification scores per example, both with K shadow principles/completions:

- **Row (fixed-query) score.** For example i with query x_i and completion y_i we form $C_i = \{c_i\} \cup$ {uniform shadow principles}. We compute $z_j = s(y_i \mid x_i, c_j)$ and report $I_{\text{row}}^{\text{diag}} = \log(K + 1) + \log \text{softmax}_j z_j$ at $j = \text{index}(c_i)$. This quantity is an InfoNCE-style score with ceiling $\log(K + 1)$. It is a valid lower bound on $I(Y; C \mid X = x_i)$ only if shadows are sampled from the true conditional marginal $p(c \mid x_i)$. With our uniform shadowing over the positive pool, we treat it as a diagnostic contrastive score, not a calibrated MI estimate (see Poole et al., 2019; CPC).

- **Column (principle $\rightarrow$ completion) score.** Holding (x_i, c_i) fixed, we assemble $Y_i = \{y_i\} \cup$ {cross-query shadow completions} and compute $I_{\text{col}}^{\text{diag}} = \log(K+1) + \log \text{softmax}_i z_i$ at $i = \text{index}(y_i)$. With exact conditional shadowing, this is again an MI bound; with uniform shadowing, it is a diagnostic that gives higher scores to principles that better select their associated completions.

A.4.3 "Clean" vs. "ungated" metrics

"Clean" always means (a) strict XML format and (b) below an entropy quantile. Unless we write "ungated," diagnostics are computed on the clean subset. Auxiliary (training) loss (cross-query, symmetric)

The auxiliary uses a cross-query score matrix $L_{ij} = s(y_i \mid x_j, c_j)$ within each micro-batch. The symmetric contrastive loss is $L_{\text{aux}} = \lambda_{\text{row}} \cdot \text{CE}(\text{softmax}_{\text{row}}(L), \text{diag}) + \lambda_{\text{col}} \cdot \text{CE}(\text{softmax}_{\text{col}}(L), \text{diag})$, with $\lambda_{\text{row}}, \lambda_{\text{col}}$ annealed as described. This auxiliary is used as a shaping regulariser; we do **not** interpret its value as a mutual information estimate due to known bias/variance trade-offs in variational MI at large MI.

We additionally include a light shaping term on the diagonal PMI-like statistic (aux-only; Equation (EqC-11)), gated by a sequence-entropy quantile, to accelerate separation without dominating GRPO.

A.5 Row-wise MI reward channel (tie-breaker)

At reward time we add a small, continuous row-wise reward derived from the row log-softmax at the positive principle with K=2 shadow principles (same sampling as A.2 Row). The mapping is a sigmoid (slope 2.5) scaled to a fixed channel weight 0.15 and controlled by an EMA autoscaler that targets ~20% of the total reward magnitude on average. Two gates apply:

- *Format gate:* only completions that exactly match the XML format (<reasoning>...</reasoning><answer>... receive the MI reward after approximately 30% of MI warmup;

- *Entropy gate:* only rows below the 80th percentile of sequence entropy receive the MI reward.

A.6   Gating and stability

- **GRPO** with **dr_grpo** loss, sequence-level importance weights, group-relative advantage centering/scaling, **epsilon=0.1**, **beta=0.0** (no KL), mask_truncated_completions=True.

- **Generations per prompt: 4** completions per prompt; temperature **1.0**, top-p **0.95**, top-k **64**, repetition penalty **1.1**.

- **MI warmup:** linearly ramp λ_{SAMI} over **50** steps; **row/col mix** anneal over **10%** of max steps from **(0.7, 0.3)→(0.5, 0.5)**.

- **Gates:** entropy gate at **80th percentile**; format gate activates after the first **30%** of MI warmup.

- **LoRA:** r=16, $\alpha = 32$, dropout **0.05**; target standard projection modules.

- **Reference policy for diagnostics:** *same model* with adapters disabled.

- **EOS & template:** Gemma-3 IT chat template with EOS=<end_of_turn>; left padding.

- These settings are sufficient for stability. Notably, **do not** add a KL penalty (**beta remains 0**), and **do not** change the group size or MI weighting unless you also retune the autoscaler.

A.7   Sinkhorn OT regulariser

We regularise hidden representations by computing an entropic Sinkhorn divergence between L2-normalised per-sequence means of last-layer hidden states on completion tokens from the current policy and the adapter-disabled reference, with warmup 200 steps and weight 0.01:

- **Blur / scaling:** 0.12 / 0.8 (GeomLoss/Sinkhorn defaults otherwise).

- **Batch subsampling:** up to 512 sequences for the OT core to keep variance modest.

A.8   Geometry probes & units

A.9   Last-token distribution drift

We log the following between current and reference last-token distributions p, q (natural-log units unless noted):

- Bhattacharyya coefficient $BC(p, q) = \sum\_k \sqrt{p\_k q\_k}$.

- Bhattacharyya angle (a.k.a. statistical angle) $\Delta_{\text{Bhat}} = \arccos\left(BC(p, q)\right)$.

- Bhattacharyya distance $D\_B = -\log BC(p,q)$ (what our plots call "Bhattacharyya distance").

- Hellinger distance $H(p,q) = \sqrt{1 - BC(p,q)}$.

- Jensen–Shannon divergence $JS(p\|q) = \frac{1}{2}\mathrm{KL}(p\|m) + \frac{1}{2}\mathrm{KL}(q\|m)$, where $m = \frac{1}{2}(p+q)$ and all quantities are reported in nats.

Note: On the categorical simplex with the Fisher metric, the closed-form Fisher–Rao geodesic distance is $d\_\mathrm{FR}(p,q) = 2\arccos\left(BC(p,q)\right) = 2\,\Delta_{\mathrm{Bhat}}(p,q)$. We sometimes display $D\_B = -\log BC(p,q)$ for numerical stability, but when discussing angles/geodesics we report Δ_{Bhat} or d_FR explicitly.

A.10 OT (offline diagnostic in our plots)

For offline analysis only, we also report an OT-style diagnostic over last-token output distributions using a token-index ground cost on the union of top-K supports (K=4096). This output-space diagnostic is separate from, and not used by, the representation-space Sinkhorn regulariser above (which operates on hidden-state summaries with Euclidean ground cost).

A.11 Reproducibility toggles

- **Dataset:** KAIST CoT-Collection train split; we take **20k** rows (length-filtered to max prompt **256**, max completion **256**).

- **Constitutions:** exactly one **positive** principle (YAML pool) per example; we compare the low-SI and high-SI sets in the paper.

- **Seeds & hardware:** single-node NVIDIA A10g (24 GB). Fix global/Torch RNG seeds if you wish to reduce variance; we observed standard seed variability around the means reported.

A.12 Additional notes

- **Fisher preconditioning.** Fisher preconditioning (diagonal proxy). When enabled, we weight token log-probs by $w_t \propto p_t(1 - p_t)$ (normalised over completion tokens). This is a practical diagonal approximation to the multi-class Fisher information (whose exact form is $\mathrm{diag}(p) - pp^\top$ for softmax logits); we use it only for the auxiliary critic when enabled.

- **LoRA reference.** We treat "reference" as "same model with adapters disabled," avoiding extra memory while giving a meaningful stationary baseline for geometry and Sinkhorn.

- **Reference KL (beta)**. We log geometry versus the adapter-disabled reference but set KL weight to zero; turning on a small KL may further stabilise large-scale runs.

B CoT/Constitution Evaluation Prompts

Note: the 'negative constitution' examples provided here were generated and used as synthetic data to validate our methods and hypotheses. They do not represent any editorial view of the ABC and exist for the validation of technical methods only.

B.1 CoT/XML Format prompt:

You will be asked a question. Respond only with your reasoning and answer between the tags:

<reasoning>{your step-by-step reasoning}</reasoning>

<answer>{concise answer}</answer>

B.1.1 Low SI constitutions – high lexical overlap

Generation prompt:

For this set of {positive principles}, generate a corresponding set of negatives.

Positive principles

- Maintain the independence and integrity of the Australian Broadcasting Corporation (ABC).

- Exercise ABC editorial control over the content the corporation broadcasts or publishes.

- Ensure that editorial decisions are not improperly influenced by political, sectional, commercial, or personal interests.

- Make reasonable efforts to ensure that material facts are accurate and presented in context.

- Do not present factual content in a way that will materially mislead the audience.

- Gather and present news and information with due impartiality.

- Present a diversity of perspectives within a reasonable timeframe.

- Do not misrepresent any perspective.

- Do not unduly favour one perspective over another.

Negative principles

- Compromise the independence and integrity for external gain.

- Yield editorial control to outside interests.

- Allow editorial decisions to be swayed by political, commercial, or personal bias.

- Present information without verifying facts or considering context.

- Present factual content in a way that deliberately misleads the audience.

- Gather and present news and information with clear bias.

- Present only a single perspective, ignoring diversity.

- Intentionally misrepresent perspectives.

- Unduly favour one perspective over others.

B.1.2 High SI Constitutions – low lexical overlap/contain procedural intent

Generation prompt:

For this set of {positive principles}, generate a set of negatives that have low lexical overlap and that emphasise procedural intent over simple negation.

Positive principles

- Maintain the independence and integrity of the Australian Broadcasting Corporation (ABC).

- Exercise ABC editorial control over the content the corporation broadcasts or publishes.

- Ensure that editorial decisions are not improperly influenced by political, sectional, commercial, or personal interests.

- Make reasonable efforts to ensure that material facts are accurate and presented in context.

- Do not present factual content in a way that will materially mislead the audience.

- Gather and present news and information with due impartiality.

- Present a diversity of perspectives within a reasonable timeframe.

- Do not misrepresent any perspective.

- Do not unduly favour one perspective over another.

Negative principles

- Publish uncorroborated claims; skip reaching out to primary sources for confirmation.

- Cite anonymous social media posts as sufficient evidence without independent verification.

- Use a single expert opinion to represent the state of the field while omitting contradictory expert views.

- Select quotes that support a predetermined angle and exclude context that would soften the claim.

- Summarise third-party reports without reading the original documents or datasets.

- Paraphrase technical results without numbers and avoid linking to underlying methodology.

- Lead with sensational framing and delay factual qualifiers until the end of the piece.

- Structure the story to present rebuttals after audience attention is likely to drop.

- Let sponsors or partners pre-approve copy changes that affect how they are portrayed.

C   Equations

C.1   Notation and conventions

Let x denote a prompt, c a (positive) principle, and y a completion. The autoregressive policy p_θ induces the sequence scores

$$S_{ij} := \log p_\theta(y_i \mid x_i, c_j), \tag{EqC-01}$$

where L_{ij} denotes a normalised score (e.g. a length- or Fisher-weighted mean). Unless noted otherwise, log denotes the natural logarithm and the subscript _2 identifies base-2 quantities. Vocabulary-level distributions are p and q.

For each row i we standardise scores via

$$\widetilde{L}_{ij} = \frac{L_{ij} - \mu_i}{\sigma_i},$$

with token weights $w_t \propto p_t(1 - p_t)$ and $\sum_t w_t = 1$.

C.2   Policy-gradient objectives (TRPO/PPO/GRPO)

Group baseline (GRPO).

$$A_i := R_i - \frac{1}{|g|} \sum_{j \in g} R_j. \tag{EqC-02}$$

PPO surrogate (sequence level).   Let $r_i = \exp(\ell_\theta - \ell_{\theta_{\mathrm{old}}})$ be the ratio of sequence log-likelihoods. The clipped surrogate to maximise is

$$\mathcal{J}_{\mathrm{PPO}} := \mathbb{E}\Big[\min\big(r_i A_i, \mathrm{clip}(r_i, 1 - \epsilon, 1 + \epsilon) A_i\big)\Big], \tag{EqC-03}$$

with the usual minimisation of $\mathcal{L}_{\mathrm{clip}} = -\mathcal{J}_{\mathrm{PPO}}$.

TRPO trust region and natural gradient.

$$\max_\theta\ \mathbb{E}[r_i A_i] \quad \text{s.t.} \quad \mathbb{E}\big[\mathrm{KL}\big(\pi_{\mathrm{old}}\big\|\pi_\theta\big)\big] \le \delta. \tag{EqC-04}$$

Locally,

$$\mathrm{KL}\big(p_\theta\big\|p_{\theta+\Delta\theta}\big) \approx \tfrac{1}{2}\Delta\theta^\top F \Delta\theta, \qquad \Delta\theta_{\mathrm{NG}} := \eta F^{-1}\nabla_\theta \mathcal{J}. \tag{EqC-05}$$

C.3   Contrastive MI (InfoNCE) and SAMI auxiliary

Let L_{ij} collect scores between completions and principle-conditioned prompts.

Row/column InfoNCE losses.

$$\mathcal{L}_{\mathrm{row}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{L_{ii}}}{\sum_{j=1}^{N} e^{L_{ij}}}. \tag{EqC-06}$$

$$\mathcal{L}_{\mathrm{col}} = -\frac{1}{N} \sum_{j=1}^{N} \log \frac{e^{L_{jj}}}{\sum_{i=1}^{N} e^{L_{ij}}}. \tag{EqC-07}$$

SAMI auxiliary (symmetric InfoNCE).

$$\mathcal{L}_{\text{SAMI}} := \lambda_{\text{row}}\mathcal{L}_{\text{row}} + \lambda_{\text{col}}\mathcal{L}_{\text{col}}. \tag{EqC-08}$$

Clean MI lower bounds (row/column). With K uniformly sampled shadow principles and the positive at $j = 0$,

$$\widehat{I}_{\text{row}}^{\text{clean}} = \log(K+1) - \widehat{\mathcal{L}}_{\text{row}}. \tag{EqC-09}$$

$$\widehat{I}_{\text{col}}^{\text{clean}} = \log(K+1) - \widehat{\mathcal{L}}_{\text{col}}. \tag{EqC-10}$$

Diagonal PMI-like statistic (diagnostics/shaping).

$$\text{diag\_MI} := \tfrac{1}{2}\left[\frac{1}{N}\sum_i \log \text{softmax}_j(L_{ij})\Big|_{j=i} + \frac{1}{N}\sum_j \log \text{softmax}_i(L_{ij})\Big|_{i=j}\right]. \tag{EqC-11}$$

C.4 MI-based reward (tie-breaker) and gating

Let $z_i = \log \text{softmax}_j \widetilde{L}_{ij}\big|_{j=0}$ (row log-probability of the positive principle among K shadows). The dense MI channel is

$$r_i^{\text{MI}} := g_i\,\beta_t\,\sigma(\alpha, z_i), \qquad g_i = \mathbf{1}\{H_i \le Q_q(H)\}\,\mathbf{1}\{\text{XML\_valid}(y_i)\}, \tag{EqC-12}$$

with an EMA autoscaler $\rho_t = \text{EMA}(|r^{\text{MI}}|)/\text{EMA}(|r^{\text{base}}|)$ and $\beta_{t+1} = \beta_t \exp\big(\eta(\rho^\star - \rho_t)\big)$.

C.5 Optimal transport and Sinkhorn divergence

Wasserstein-2 geometry.

$$W_2^2(\alpha, \beta) := \min_{\pi \in \Pi(\alpha,\beta)} \sum_{a,b} \pi_{ab}\,\|x_a - y_b\|_2^2. \tag{EqC-13}$$

Entropic OT (Cuturi).

$$\text{OT}_\varepsilon(\alpha, \beta) := \min_{\pi \in \Pi(\alpha,\beta)} \sum_{a,b} \pi_{ab}C_{ab} + \varepsilon\,\text{KL}\big(\pi\big\|\alpha \otimes \beta\big), \tag{EqC-14}$$

with $C_{ab} = \|x_a - y_b\|_2^2$.

Sinkhorn divergence (debiased entropic OT).

$$S_\varepsilon(\alpha, \beta) := \text{OT}_\varepsilon(\alpha, \beta) - \tfrac{1}{2}\text{OT}_\varepsilon(\alpha, \alpha) - \tfrac{1}{2}\text{OT}_\varepsilon(\beta, \beta). \tag{EqC-15}$$

Representation-space OT regulariser. Let μ_θ and μ_{ref} be empirical measures of (normalised) hidden-state summaries under the current and reference policies. Define

$$\mathcal{R}_{\text{OT}} := \lambda_{\text{OT}}\,S_\varepsilon(\mu_\theta, \mu_{\text{ref}}). \tag{EqC-16}$$

C.6 Unified ENIGMA objective

Combining policy improvement, the SAMI auxiliary, and OT regularisation:

$$\mathcal{L}_{\text{ENIGMA}} := \mathcal{L}_{\text{GRPO}} + \lambda_{\text{SAMI}} \mathcal{L}_{\text{SAMI}} + \mathcal{R}_{\text{OT}}. \tag{EqC-17}$$

C.7 Geometry probes (output and representation)

Bhattacharyya coefficient and angle (last token).

$$\text{BC}(p,q) = \sum_k \sqrt{p_k q_k}, \qquad \Delta_{\text{Bhat}} = \arccos\left(\text{BC}(p,q)\right). \tag{EqC-18}$$

Hellinger distance.

$$H(p,q) = \sqrt{1 - \text{BC}(p,q)}. \tag{EqC-19}$$

Jensen–Shannon divergence (bits).

$$\text{JS}_2(p\|q) := \tfrac{1}{2}\text{KL}_2(p\|m) + \tfrac{1}{2}\text{KL}_2(q\|m), \qquad m = \tfrac{1}{2}(p+q). \tag{EqC-20}$$

Fréchet (Gaussian) distance.

$$d_F^2 = \|\mu_1 - \mu_2\|_2^2 + \text{tr}\left(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right). \tag{EqC-21}$$

Effective rank and participation ratio. Let $\{\lambda_i\}$ be eigenvalues of a covariance and $p_i = \lambda_i / \sum_j \lambda_j$. Then

$$\text{effrank} = \exp\left(-\sum_i p_i \log p_i\right), \qquad \text{PR} = \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2}. \tag{EqC-22}$$

Linear CKA. For centred design matrices X, Y (rows correspond to examples),

$$\text{CKA}_{\text{lin}} = \frac{\|Y^\top X\|_F^2}{\|X^\top X\|_F \|Y^\top Y\|_F}. \tag{EqC-23}$$

C.8 Evaluation/diagnostic scalars

Predictive information and perplexity mapping.

$$\Delta\text{NLL} = \text{NLL}_{\neg c} - \text{NLL}_{+c}, \qquad \frac{\text{PPL}_{+c}}{\text{PPL}_{\neg c}} = 2^{-\Delta\text{NLL}}. \tag{EqC-24}$$

AUC (Mann–Whitney).

$$\text{AUC} = \Pr[s^+ > s^-] + \tfrac{1}{2}\Pr[s^+ = s^-]. \tag{EqC-25}$$

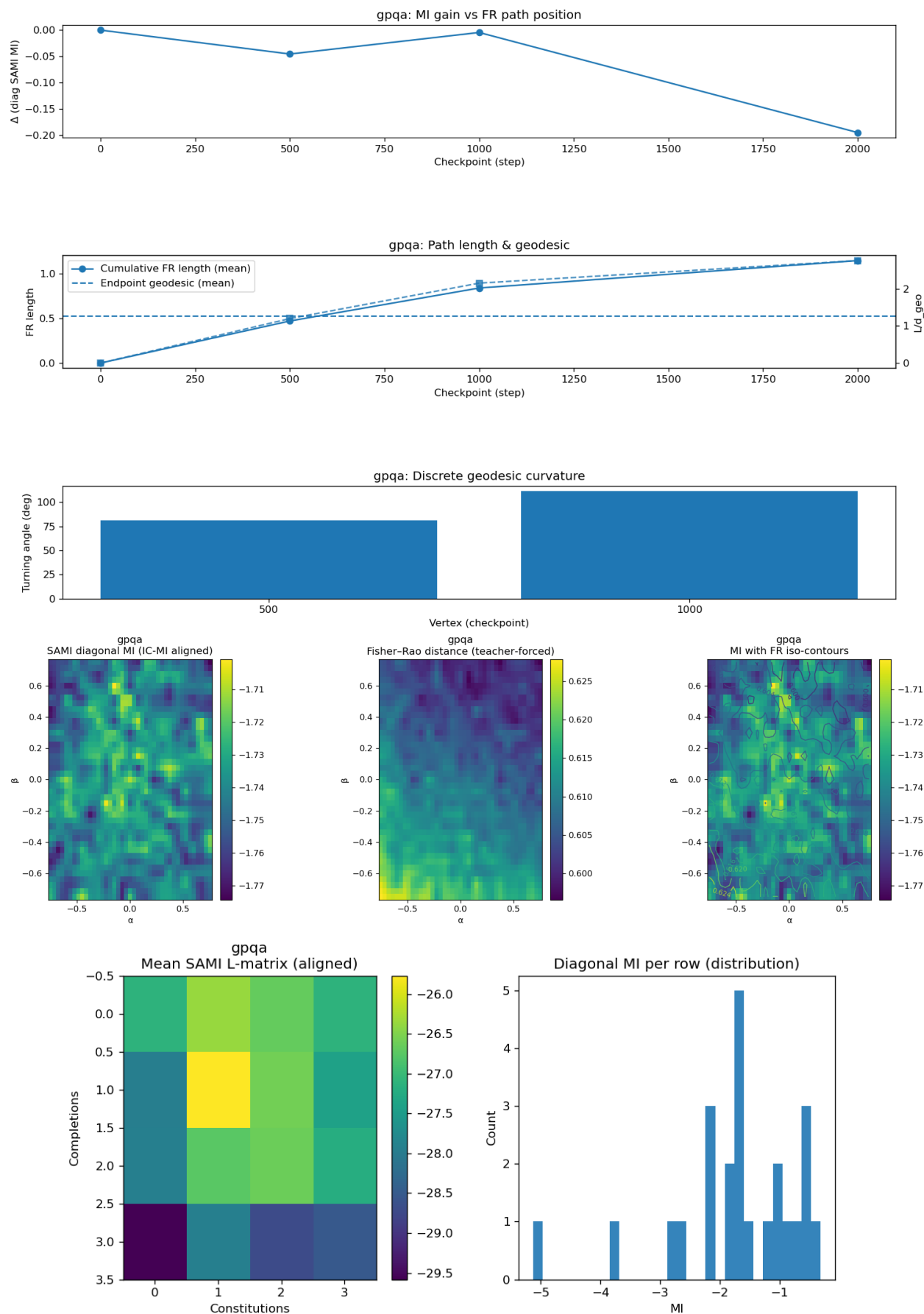**Sufficiency Index (SI).** With z-scores for bits, MI, and separation,

$$\text{SI} = w_b z_{\text{bits}} + w_m z_{\text{MI}} + w_s z_{\text{sep}}. \tag{EqC-26}$$

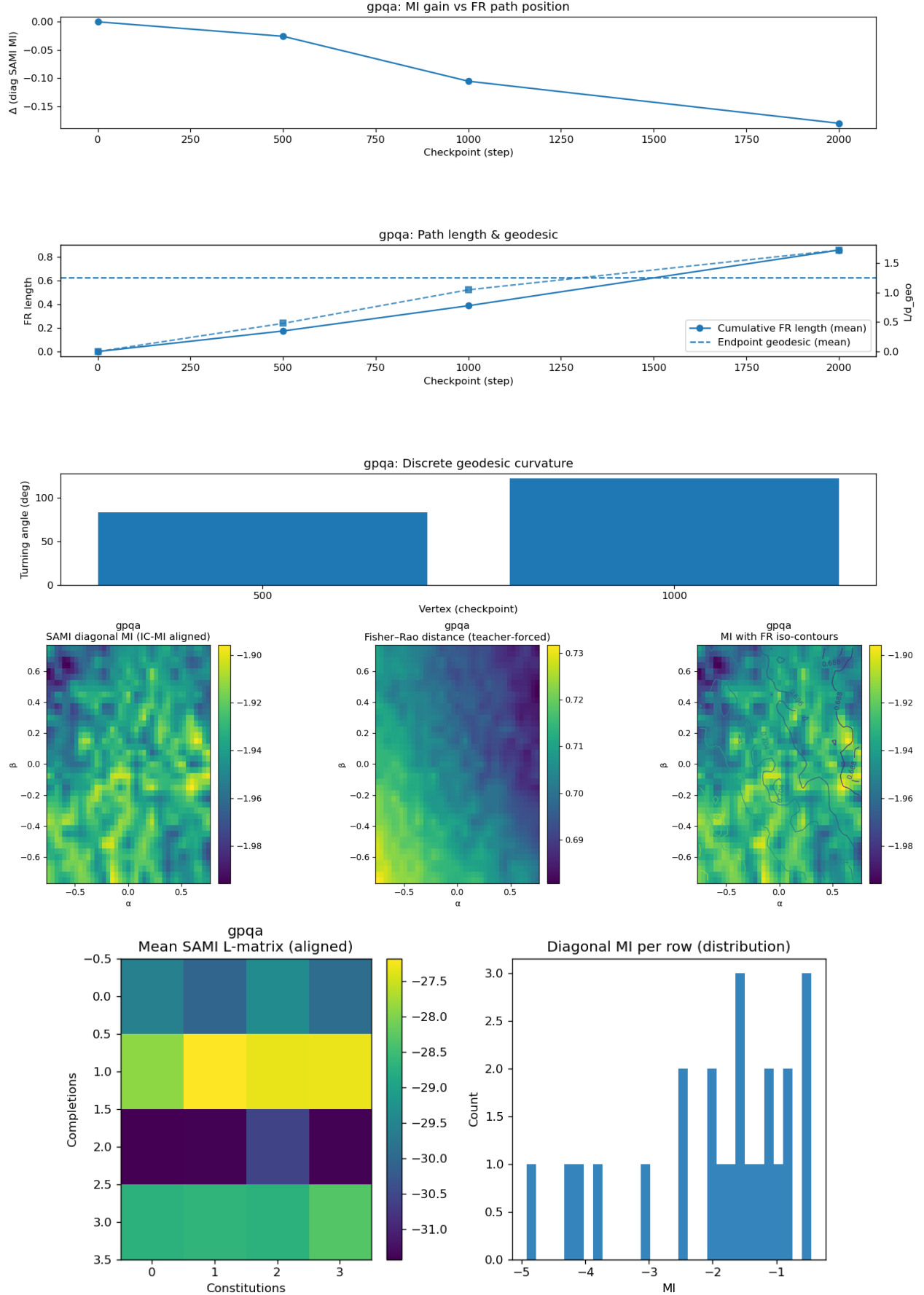## C.9 Spaces and empirical measures for hidden states

Let $\mathcal{P}_2(\mathbb{R}^d)$ denote probability measures with finite second moment. For a batch of sequence summaries $\{\bar{h}_i\}_{i=1}^B \subset \mathbb{R}^d$, define

$$\mu_\theta := \frac{1}{B} \sum_{i=1}^B \delta_{\bar{h}_i} \in \mathcal{P}_2(\mathbb{R}^d), \tag{EqC-27}$$

and analogously μ_{ref}, which is used in Equation (EqC-16).

Figure 3: Information-geometry diagnostics for ENIGMA High-SI on GPQA. Panel A tracks the SAMI MI delta, Fisher–Rao path length, and turning angles across checkpoints; Panel B overlays MI ridges with Fisher–Rao iso-contours; Panel C visualises the row-wise InfoNCE binding matrix and histogram.

Figure 4: Information-geometry diagnostics for ENIGMA Low-SI on GPQA. Panel A charts the SAMI MI delta alongside Fisher–Rao path quantities; Panel B shows the MI landscape against Fisher–Rao iso-contours; Panel C reports the row-wise InfoNCE binding heatmap and histogram.
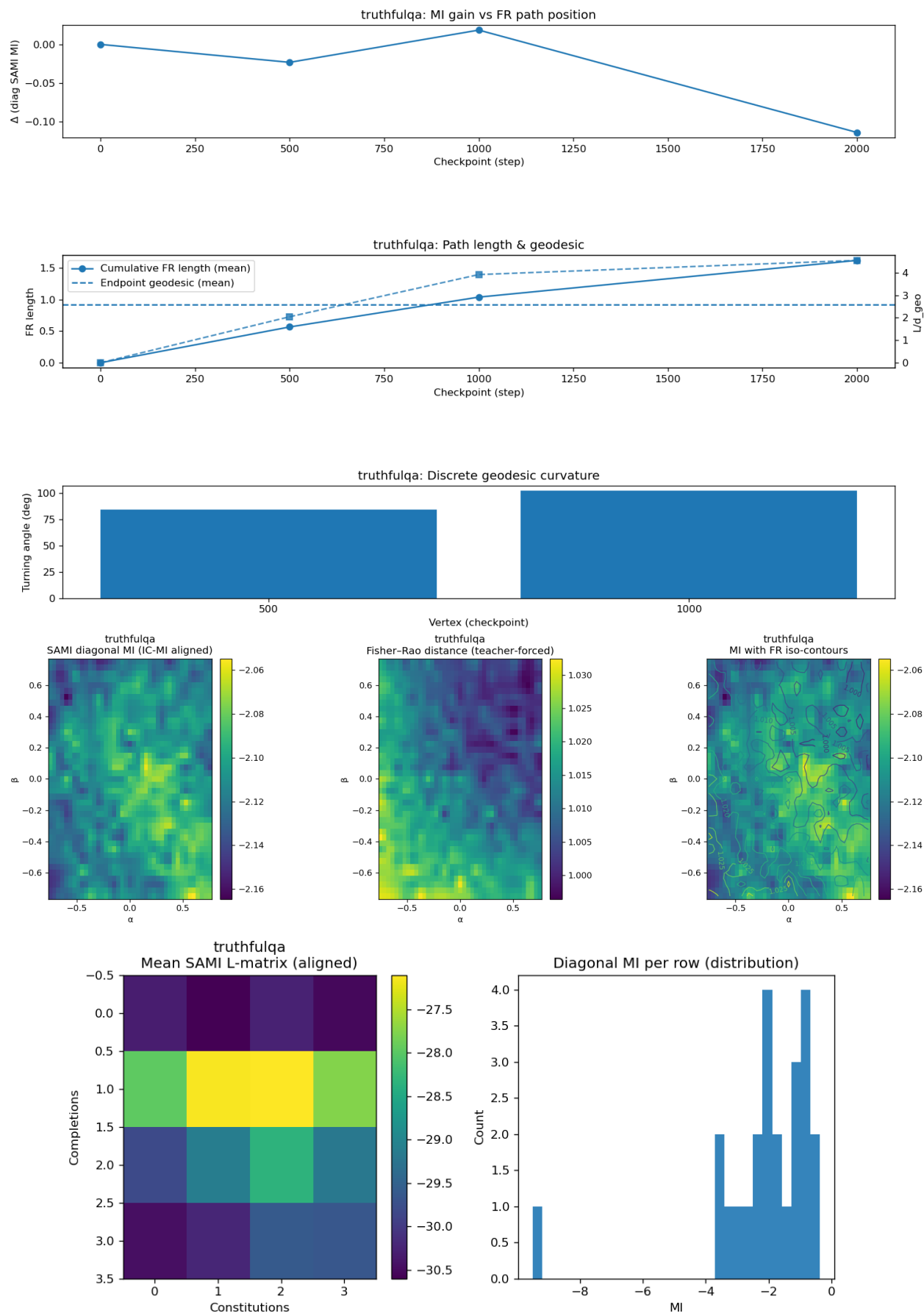
Figure 5: Information-geometry diagnostics for the GRPO CoT ablation on GPQA. Panel A records the SAMI MI delta and Fisher–Rao trajectory; Panel B plots MI landscapes with Fisher–Rao contours; Panel C shows binding structure and histograms.
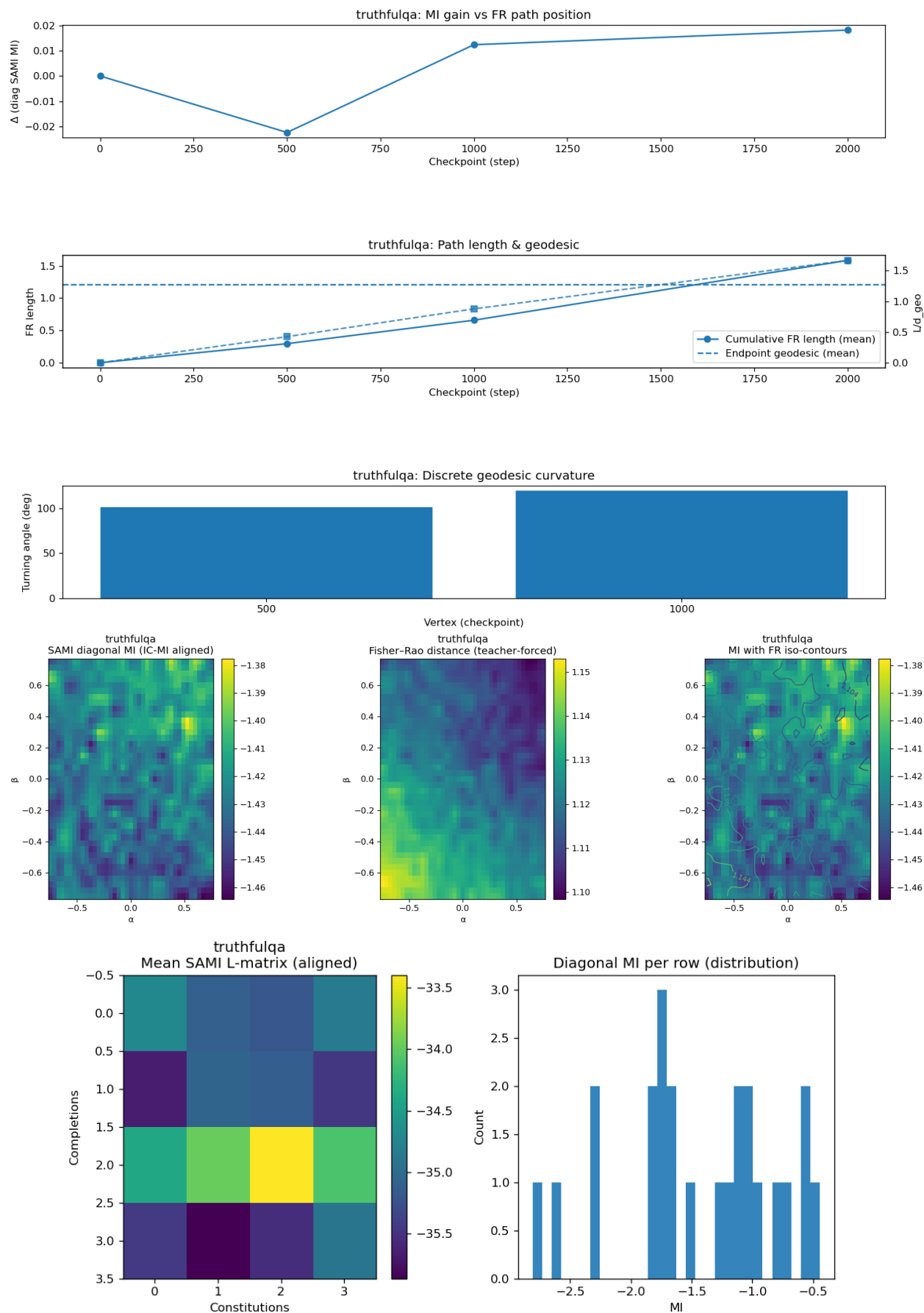
Figure 6: Information-geometry diagnostics for the GRPO CoT+ ablation on GPQA. Panel A reports the SAMI MI trajectory and Fisher–Rao path; Panel B overlays MI landscapes with Fisher–Rao contours under stochastic jitter; Panel C highlights binding patterns.
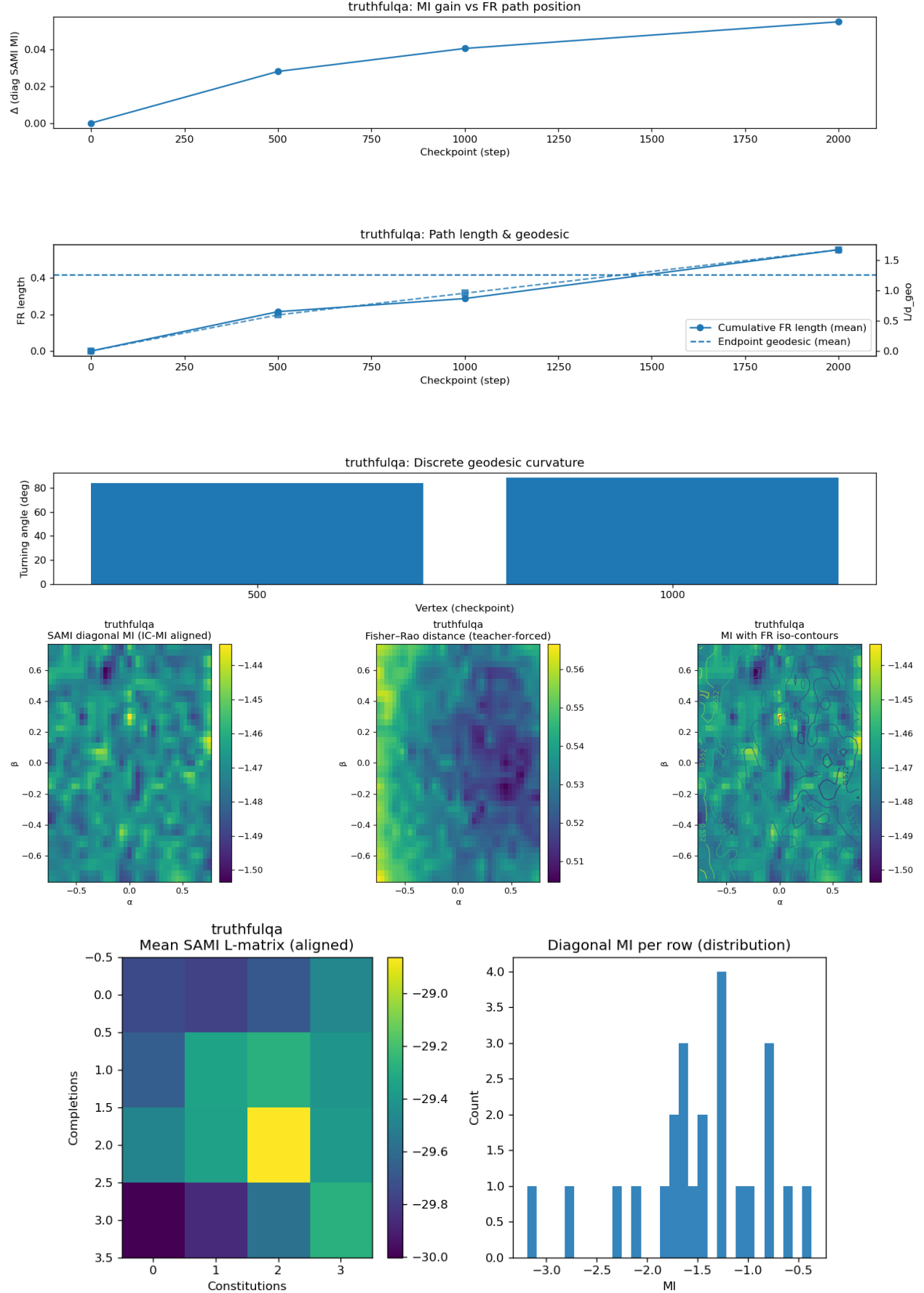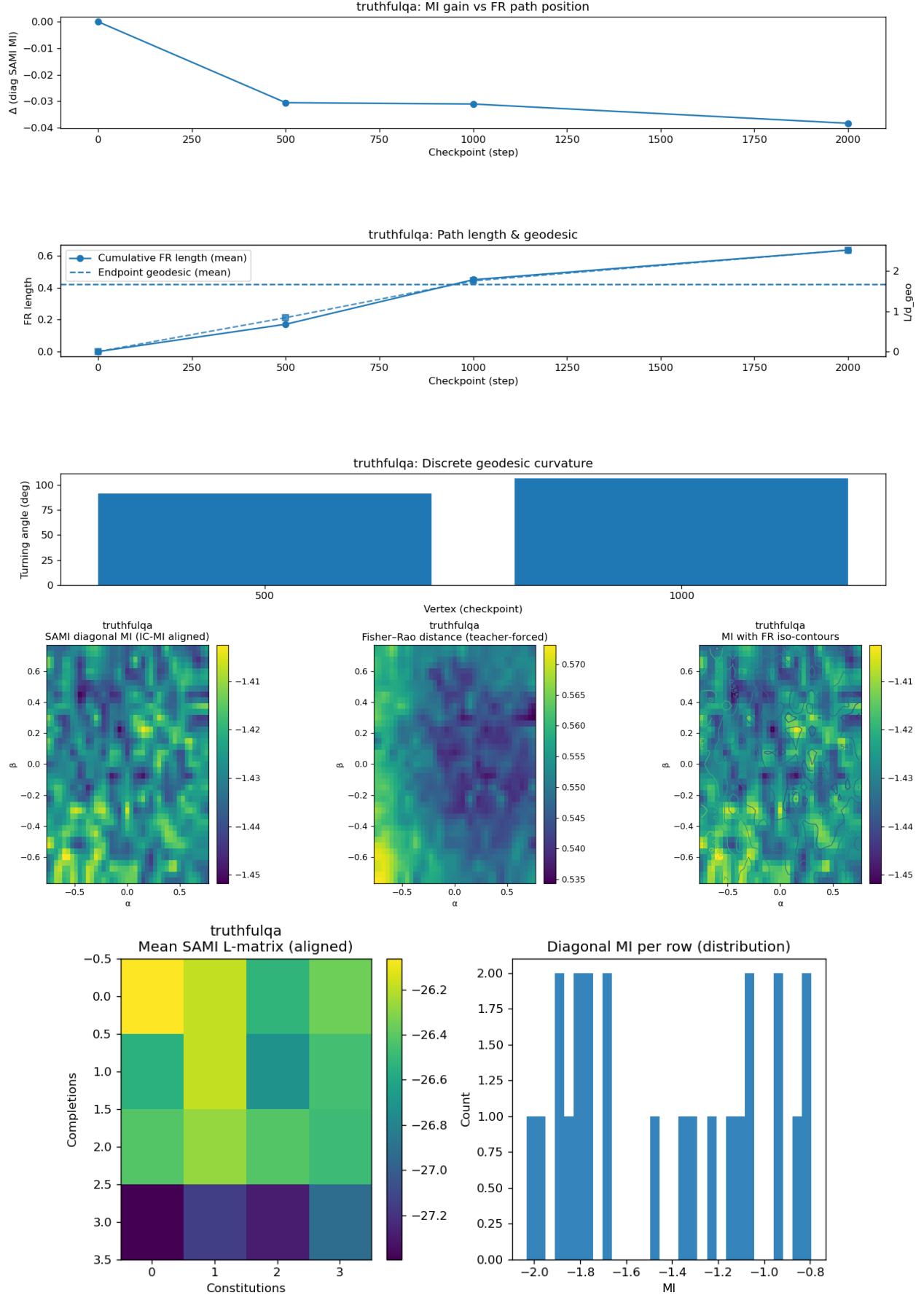
Figure 7: Information-geometry diagnostics for ENIGMA High-SI on TruthfulQA. Panel A plots the SAMI MI trajectory and Fisher–Rao path; Panel B overlays MI ridges with Fisher–Rao contours; Panel C shows the InfoNCE binding heatmap and histogram.

Figure 8: Information-geometry diagnostics for ENIGMA Low-SI on TruthfulQA. Panel A covers the SAMI MI trajectory and Fisher–Rao path; Panel B juxtaposes MI ridges with Fisher–Rao contours; Panel C shows binding matrices and histograms.

Figure 9: Information-geometry diagnostics for the GRPO CoT ablation on TruthfulQA. Panel A plots the SAMI MI delta with Fisher–Rao paths; Panel B maps MI ridges to Fisher–Rao contours; Panel C summarises binding structure.

Figure 10: Information-geometry diagnostics for the GRPO CoT+ ablation on TruthfulQA. Panel A tracks the SAMI MI delta and Fisher–Rao trajectory; Panel B overlays MI landscapes with Fisher–Rao contours; Panel C presents the InfoNCE binding view.