SEMFAITH: SEMANTIC GEOMETRY FOR CHAIN-OF-THOUGHT FAITHFULNESS EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a reasoning trace reveal when a model conceals its use of a misleading hint? We investigate this question through the geometry of sentence embeddings, using BonaFide's hint-acknowledgment labels as an observable test case. Our hypothesis is that how a trace's steps spread across semantic regions, how abruptly they shift topic, and how far they stray from reference centroids carry information about whether the hint is acknowledged. SemFaith segments a chain of thought (CoT) into steps, embeds them with a fixed encoder, and summarizes these measurements at three clustering resolutions; a regularized linear head maps the resulting 24 features to a score without accessing the audited model's hidden states or querying an auxiliary judge. Across five held-out Instruct models (662 CoTs), SemFaith attains 0.885 mean fold AUROC. On 67 within-model pairs closely matched for step and word counts, its AUROC is 0.788, compared with 0.533 for a length-only control, showing signal beyond gross trace length. A supervised TF-IDF baseline reaches 0.991 on the full held-out evaluation, exposing strong lexical cues in this benchmark. These results identify a compact geometric predictor of hint concealment whose signal is not explained by gross trace length, while delimiting what BonaFide can establish about general reasoning faithfulness.

1 INTRODUCTION

Chain-of-thought (CoT) prompting Wei et al. (2022) and reasoning language models Achiam et al. (2023); Liu et al. (2024) expose intermediate text that readers often use to inspect a model's answer. Yet a fluent account can omit a cue that influenced the answer Turpin et al. (2023); Lanham et al. (2023); Paul et al. (2024). This matters especially when auditors can observe the generated text but cannot inspect model internals or run repeated interventions. A practical question follows: can the trace itself provide a useful signal about such omissions?

BonaFide Gur-Arieh et al. (2026) provides a concrete setting. Its tasks insert misleading hints, including sycophancy cues Sharma et al. (2024), and label whether CoTs acknowledge or conceal the hint's influence. This operational definition makes one failure mode measurable at the level of whole traces. The answer can be the same while the trace differs in whether it reports the hint. A model that acknowledges a hint may use different words, produce a shorter trace, or follow a different task template from a model that conceals it. A successful text-only predictor may exploit these regularities without detecting a property shared by other kinds of unfaithful reasoning.

Existing evaluation methods approach faithfulness through distinct observations. Attribution compares signals associated with an answer and its explanation Parcalabescu & Frank (2024); interventions alter a CoT and measure answer changes Lanham et al. (2023); LLM judges read the generated explanation in context. Each can be informative, but the first requires access to model probabilities, the second measures causal dependence under an intervention, and the third requires another model. We study a complementary, inexpensive representation computed from the output trace alone. Scores obtained under different operational definitions and evaluation sets are not directly comparable, so our main comparisons use the same held-out models and folds.

Our starting hypothesis is that the distribution and ordering of step meanings may carry information about whether a hint is acknowledged. A trace can range across semantic regions, remain concentrated around a few recurring ideas, or contain a sharp transition that stands apart from its surrounding steps. Such properties are observable after mapping steps into a fixed sentence-embedding

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

space. They do not by themselves reveal a model's internal computation: the hypothesis is an empirical claim about a textual proxy, to be tested against simpler signals rather than assumed from an attractive visualization.

SemFaith makes that claim testable through four stages. It segments and embeds each CoT; builds training-fold MiniBatch K-Means reference spaces at $K \in \{5, 10, 20\}$; computes eight measurements at each resolution, covering outlier distance, local coherence, and cluster occupancy; and feeds the 24-dimensional vector to a regularized logistic regression. The encoder is fixed and the score needs neither the audited model's hidden states nor an auxiliary LLM. Multiple resolutions let the same trace be described at coarse and finer granularities, while the low-dimensional representation permits feature-family checks.

We evaluate complete held-out models, then probe two central alternative explanations. First, a length-only model is already strong (0.869 AUROC), so we additionally compare predictions on 67 within-model positive–negative pairs closely matched for step and word counts. SemFaith retains 0.788 AUROC on these pairs, compared with 0.533 for length. Second, a TF-IDF text baseline reaches 0.991 AUROC under the same model-held-out folds. This result exposes a lexical shortcut and rules out a claim that geometric features are the strongest available text representation on BonaFide. Under the same held-out-model folds, the multi-resolution geometry score (0.885) nonetheless exceeds each tested single resolution (K=5: 0.859; K=10: 0.842; K=20: 0.854); the advantage is a specific, inspectable property of this representation.

Our contribution is therefore a concrete way to represent CoT step structure and an evaluation that states both its value and its limits. The geometry captures signal beyond coarse trace length for BonaFide's hint-concealment labels. The lexical control and the limited range of labels leave broader claims about hidden computation and other faithfulness failures open. To facilitate reproducibility, our code is publicly available.[1].

2 Related Work

Faithfulness and its measurement. A generated explanation is faithful when it accurately reflects the process relevant to the prediction Jacovi & Goldberg (2020). For CoTs, this is difficult to verify from text alone. Perturbation studies truncate, replace, or edit reasoning and measure answer changes Lanham et al. (2023). These interventions reveal whether content is useful to the model under a modified prompt, which can differ from whether the original trace truthfully reported an earlier influence. BonaFide Gur-Arieh et al. (2026) focuses on that distinction by placing misleading hints in tasks and checking whether the model admits using them. Its label is unusually concrete, but it represents one operationalization of faithfulness.

Controlled causal evaluations sharpen this distinction: Zaman and Srivastava Zaman & Srivastava (2025) construct model-edited faithful/unfaithful explanation pairs, and Matton et al. Matton et al. (2025) compare explanation-implied importance with measured causal effects. Such designs test whether a metric reflects the causal role of an explanation; SemFaith instead predicts the observable hint-concealment label, and aligning labels across these settings remains future work.

Attribution and judge methods. CC-SHAP Parcalabescu & Frank (2024) compares feature attributions for predictions and explanations, drawing on Shapley-style explanations Lundberg & Lee (2017); related LLM attribution methods study alternative approximations and interactions Naudot et al. (2025); Fumagalli et al. (2026); Walker & Ewetz (2025). They require probability or internal access and repeated model evaluation. LLM judges can read an entire reasoning trace and assess its plausibility or consistency, but judge reliability itself is an active research topic Mittal & Arike (2026). BonaFide reports metrics from these families under its own protocols. We describe them as context, while reserving numerical head-to-head claims for predictors evaluated on identical examples and held-out models. RFEval Han et al. (2026) and FaithCoT-Bench Shen et al. (2025) use different evaluation designs; success on hint-concealment labels does not establish performance on those tasks.

Representations of generated text. Sentence-BERT Reimers & Gurevych (2019) and clustering Sculley (2010) allow step sequences to be summarized without querying the source model.

[1] https://anonymous.4open.science/r/SemFaith/

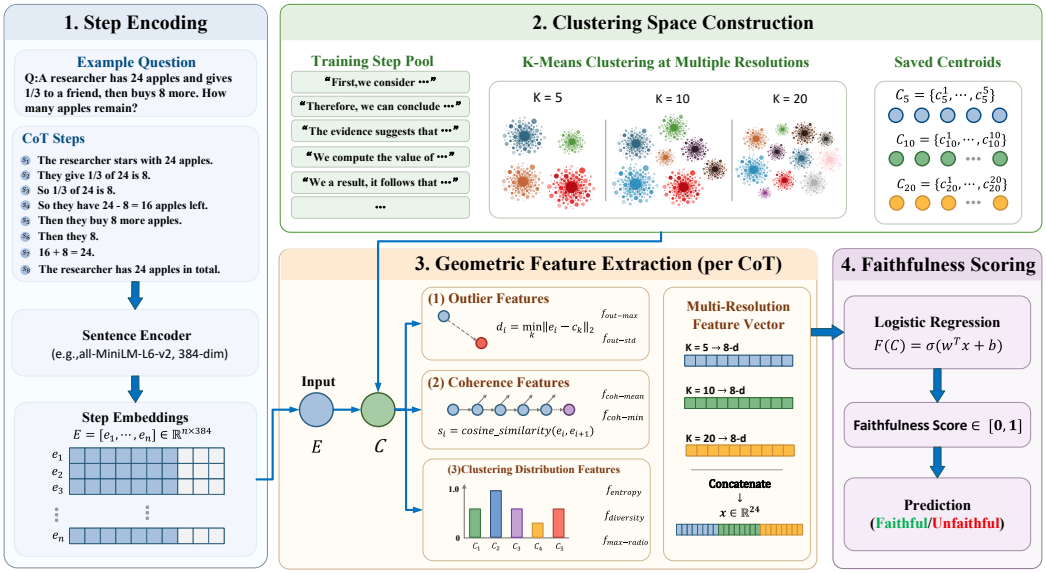

Figure 1: The four-stage SemFaith framework: step encoding, training-fold reference clusters at three resolutions, geometric summaries, and a regularized faithfulness score. Centroids and the scoring head are fitted before evaluating a held-out model.

Embedding geometry and its distortions have been studied in neighboring settings Tehenan (2025); Nielsen & Hansen (2024); Marks & Tegmark (2024); geometric summaries are also used for uncertainty or inconsistency in generation Kuhn et al. (2023); Manakul et al. (2023). SemFaith applies them to the distribution and succession of CoT steps. Entropy of cluster occupancy, distance to a centroid, and adjacent-step similarity are properties of observed text, not direct observations of reasoning mechanisms. We test these features against step and word counts and a supervised lexical representation. The latter is especially relevant because acknowledging a hint may leave direct lexical evidence in a trace.

These comparisons position SemFaith as an interpretable structural predictor. Its 24 inputs describe the trace through a small set of measurements; TF-IDF describes a much larger vocabulary of surface cues. Their relative performance helps separate the existence of a geometric signal from any claim that geometry alone identifies why an answer was produced.

3 Method

3.1 Problem Formulation

Let q be a natural-language question. An LLM generates a CoT $C = [s_1, s_2, \ldots, s_n]$ with n reasoning steps and a final answer a. Ground-truth faithfulness labels $y \in \{0, 1\}$ are obtained through BonaFide's task-design methodology Gur-Arieh et al. (2026): a misleading hint is embedded in the question such that if the model produces the hinted answer, the hint can be shown to have influenced the answer. The CoT is labeled faithful ($y = 1$) if it honestly acknowledges hint influence, and unfaithful ($y = 0$) if it fabricates an independent justification while concealing the hint. The labeling pipeline achieves 98.9% precision against human annotation.

We learn a text-only scoring function $\mathcal{F} : C \mapsto [0, 1]$, evaluated by its ranking quality (AUROC) against ground-truth labels. SemFaith implements $\mathcal{F}$ through four stages (Figure 1), detailed below.

3.2 Step Encoding

CoT text is segmented using a multi-stage heuristic tailored to reasoning text. We split by newlines (the natural organization of CoT outputs), merge short fragments to avoid pseudo-steps, split long

blocks at reasoning transition markers (e.g., *Therefore*, *However*, *First*), and filter fragments below a minimum length.

Each step s_i is encoded via a pre-trained sentence encoder:

$$\mathbf{e}_i = \text{Enc}(s_i) \in \mathbb{R}^{384}, \quad i = 1, \ldots, n. \tag{1}$$

For a CoT with n steps, the embedding matrix is $E = [\mathbf{e}_1, \ldots, \mathbf{e}_n] \in \mathbb{R}^{n \times 384}$. We use the fixed 384-dimensional all-MiniLM-L6-v2 encoder Reimers & Gurevych (2019). Encoding is shared by all three clustering resolutions; the 662 evaluated CoTs contain 7,413 segmented steps.

3.3 Clustering Space Construction

To capture prototypical reasoning patterns at multiple semantic granularities, we build K-Means clustering spaces at three resolutions $K \in \{5, 10, 20\}$. Within each training fold, we randomly sample up to $M = 4{,}000$ step embeddings as the reference set $\mathcal{R}$; for each K, MiniBatchKMeans Sculley (2010) is fit on $\mathcal{R}$ with 3 random seeds, and the resulting features are averaged across seeds to reduce initialization variance. The three 8-dimensional blocks are concatenated into a 24-dimensional representation. The three resolutions supply coarse, intermediate, and finer occupancy and distance descriptions. We do not assign fixed cognitive roles to their centroids: clustering is refitted within each training fold and the centroids are descriptive reference points.

3.4 Geometric Feature Extraction

For each CoT with embedding matrix E, we extract eight geometric features across three dimensions. These features operationalize the core geometric hypothesis: different traces can exhibit dispersed or concentrated cluster occupancy, variable distance from reference centroids, and different degrees of local continuity. We test whether those measurements predict BonaFide labels.

3.4.1 Outlier Features

Each step's Euclidean distance to its nearest centroid summarizes its position relative to the fitted reference space: $d_i = \min_k \|\mathbf{e}_i - \mathbf{c}_k\|_2$. Three aggregations capture complementary aspects of distributional deviation:

$$f_{\text{out-mean}} = \frac{1}{n} \sum_{i=1}^{n} d_i, \tag{2}$$

$$f_{\text{out-max}} = \max_i d_i, \tag{3}$$

$$f_{\text{out-std}} = \text{std}(d_1, \ldots, d_n). \tag{4}$$

$f_{\text{out-max}}$ records the most distant step; $f_{\text{out-std}}$ records how uneven those distances are. These statistics can expose a local deviation that a chain average would conceal, although distance alone cannot identify whether a step is fabricated.

3.4.2 Coherence Features

Adjacent-step cosine similarity measures the semantic continuity of reasoning Barzilay & Lapata (2008):

$$s_i = \frac{\langle \mathbf{e}_i, \mathbf{e}_{i+1} \rangle}{\|\mathbf{e}_i\| \cdot \|\mathbf{e}_{i+1}\|}, \quad i = 1, \ldots, n-1. \tag{5}$$

Two features summarize the coherence profile:

$$f_{\text{coh-mean}} = \frac{1}{n-1} \sum_{i=1}^{n-1} s_i, \tag{6}$$

$$f_{\text{coh-min}} = \min_i s_i. \tag{7}$$

$f_{\text{coh-min}}$ captures the single most severe semantic break in the reasoning chain. The minimum can reveal a sharp semantic transition that an average can obscure. Neither statistic determines whether that transition was honest or misleading.

3.4.3 CLUSTERING DISTRIBUTION FEATURES

Step-to-centroid assignments $\ell_i = \arg\min_k \|\mathbf{e}_i - \mathbf{c}_k\|$ produce cluster occupancy counts cnt_k and proportions $p_k = \text{cnt}_k/n$. Three features quantify the distribution of steps among fitted clusters:

$$f_{\text{entropy}} = -\frac{1}{\log K} \sum_{k=1}^{K} p_k \log(p_k + \varepsilon), \tag{8}$$

$$f_{\text{diversity}} = \frac{|\{k : \text{cnt}_k > 0\}|}{K}, \tag{9}$$

$$f_{\text{max-ratio}} = \max_k p_k. \tag{10}$$

f_{entropy} is normalized via division by $\log K$ Shannon (1948); Rosenberg & Hirschberg (2007), with $\varepsilon = 10^{-8}$ smoothing matching the released code: a value near 0 indicates one occupied cluster, and the uniform distribution over K clusters yields a value approaching 1. Its attainable maximum is bounded by $\log \min(n, K)/\log K$, so short traces cannot reach 1 when $n < K$. $f_{\text{diversity}}$ records the occupied fraction of clusters and $f_{\text{max-ratio}}$ records the largest occupancy share. These measurements are also sensitive to the number of steps; we therefore compare them with length controls rather than interpreting concentration as a direct test of honesty.

3.5 SCORING

For the single-resolution setting, the feature vector is $\mathbf{x} \in \mathbb{R}^8$; for the multi-resolution ensemble, we concatenate the three 8-dimensional blocks into $\mathbf{x} \in \mathbb{R}^{24}$. We Z-score normalize each feature using training-fold statistics and feed the result into L2-regularized logistic regression:

$$\mathcal{F}(C) = \sigma(\mathbf{w}^\top \mathbf{x}_{\text{norm}} + b), \quad \sigma(z) = \frac{1}{1 + e^{-z}}. \tag{11}$$

We set inverse regularization strength $C = 0.1$ and use `class_weight='balanced'` for the 168 acknowledged and 494 concealed CoTs in the five-model analysis. The 24 coefficients and intercept define a small, inspectable scoring head. Because features are correlated, individual coefficients should not be read as independent causal contributions; our feature-family ablations provide a more direct predictive check.

4 EXPERIMENTS

4.1 DATASET, EVALUATION, AND CONTROLS

Dataset and target. BonaFide Gur-Arieh et al. (2026) includes 1,120 CoT-level records across ten models, together with step-level annotations and several hint constructions. The original SemFaith evaluation described 762 traces as an Instruct-only subset; inspection of the model identifiers shows that this set also includes 100 DeepSeek Distill traces with no positive CoT labels. We therefore use the five models explicitly named Instruct as the primary evaluation set: 662 CoTs, including 168 hint acknowledgments and 494 concealments. The full dataset remains useful for embedding-cache construction and for reproducing the prior result, but it is not mixed into the new five-model comparison. The label denotes acknowledgement of a successful misleading hint, not a direct measurement of hidden computation.

Held-out-model protocol. We use five Group K-Fold splits by model Pedregosa et al. (2011): each split tests on all traces of one Instruct model and trains on the other four. The sentence encoder is fixed in advance. For each fold, the reference steps, MiniBatch K-Means centroids, feature standardization, and logistic-regression coefficients are fitted using training models only. We use $K = 5, 10, 20$, the same three seeds for each resolution, up to 4,000 training steps per fit, and an L2 logistic head with $C = 0.1$ and balanced class weights. Feature values are averaged over seeds before fitting the head; the primary number is the unweighted mean of five held-out-model AUROCs. We give model-specific scores in the supplementary appendix; the model, not an individual trace, is the generalization unit.

Controls. We fit three predictors on the same folds. The length control uses segmented-step count, word count, and words per step. The lexical control uses up to 5,000 TF-IDF unigram and bigram

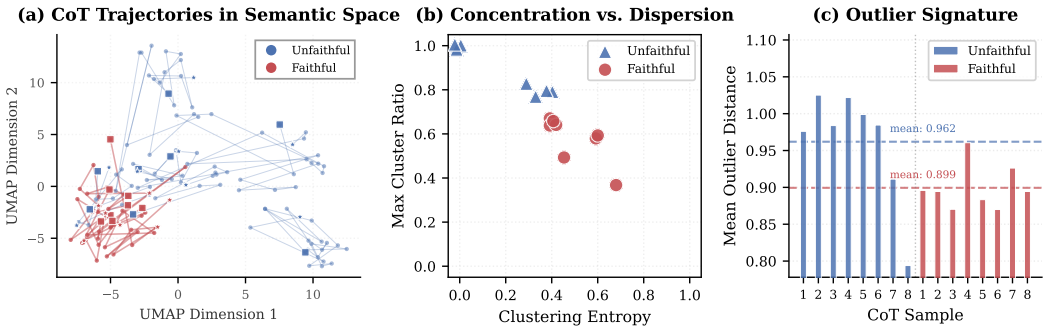

Figure 2: Illustrative geometry from the original visualization. Eight CoTs per class were selected to display clear signatures. (a) UMAP projection of their step embeddings and trajectories; (b) cluster occupancy entropy versus maximum occupancy share; (c) mean distance to fitted centroids. The examples explain the measurements but are not used as evidence that the full dataset separates this cleanly.

Table 1: CoT-level mean AUROC across five held-out Instruct models (662 traces). All rows use the same folds and training-only fitting. These are unweighted means of model-level AUROCs.

Predictor	AUROC-CoT
Length: steps, words, words/step	0.869
Geometry, $K = 5$	0.859
Geometry, $K = 10$	0.842
Geometry, $K = 20$	0.854
SemFaith, $K = 5, 10, 20$	**0.885**
SemFaith + length	0.912
TF-IDF text	0.991

features fitted on training-fold CoT text. Both use the same class weighting and logistic-regression regularization as SemFaith. These controls test distinct alternatives: whether geometry is a proxy for how much the model wrote, and whether direct wording is a stronger shortcut. We also concatenate length and geometry features to see whether their information overlaps. No label from the held-out model is used to fit a control.

4.2 Geometric view of the traces

Figure 2 retains the original trajectory visualization to explain what the features measure. It shows eight examples per class selected for clear contrast, projected with UMAP. The colored paths and the accompanying entropy and outlier panels illustrate possible geometric signatures; selection and two-dimensional projection prevent this figure from serving as independent evidence of class-wide separation. The full held-out sample appears in Figure 3, where occupancy and length patterns can be examined without selecting favorable traces. The distinction is useful: a visualized example motivates a representation, whereas quantitative held-out evaluation tests it.

4.3 Held-out-model prediction

Table 1 reports every predictor under the same five held-out-model splits. SemFaith reaches 0.885 mean fold AUROC. The strongest single-resolution variant, $K = 5$, reaches 0.859, so using three resolutions adds 0.026 under this protocol. This supports retaining multiple granularities in the representation. The gain is modest and does not imply that each resolution adds unique information on every held-out model; the per-model scores vary, as the supplementary appendix shows.

The length control itself reaches 0.869, only 0.016 below SemFaith, so the full-set AUROC cannot isolate geometric structure from verbosity. Combining geometry with explicit length raises AUROC to 0.912, consistent with partially complementary predictive information, but still below the lexical

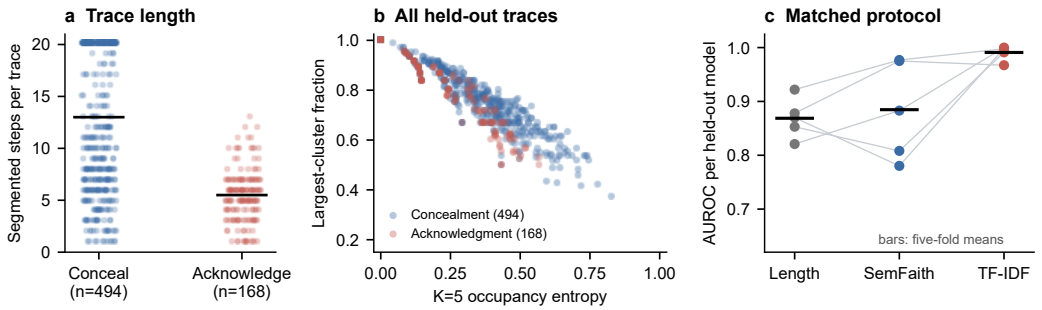

Figure 3: Instruct-model evaluation. (a) Step counts for all 662 CoTs by label, with median bars. (b) $K = 5$ occupancy entropy and largest-cluster fraction for every trace; class distributions overlap, and density is greatest along a length-dependent pattern. (c) AUROC for each held-out model under identical splits, with black bars indicating five-model means.

model. TF-IDF reaches 0.991 on the same held-out models. This is an unusually strong text-only result and substantially changes the interpretation of the benchmark: wording associated with admitting or concealing the hint remains predictive across models. The geometric approach offers a low-dimensional structural view; it is not the best supervised text predictor in this setting.

Published BonaFide leaderboard values for SKY Judge, CC-SHAP, perturbation metrics, and other methods provide context for the cost and access requirements of alternative approaches Gur-Arieh et al. (2026). They were not rerun on the 662 traces and identical folds used here. Consequently, we do not rank SemFaith above or below those methods numerically. The comparison supported by Table 1 is among our predictors under a common protocol.

4.4 Length restriction and full-sample diagnostics

In the evaluated set, acknowledged traces average 5.45 segmented steps and 141 words, whereas concealed traces average 13.15 steps and 764 words. This pronounced imbalance makes matching necessary. Within each held-out model, we pair acknowledged and concealed traces without replacement, minimizing Euclidean distance in $\log(1 + \text{steps})$ and $\log(1 + \text{words})$ and retaining only pairs within a 0.25 caliper. The resulting 67 pairs have nearly equal mean steps (6.10 versus 6.12) and much closer mean words (137 versus 147). We score those traces with the predictors already trained on the full training folds; no model is refitted on the selected test pairs.

On the matched set, length-only AUROC falls to 0.533 while SemFaith retains 0.788. This is evidence that its ranking signal is not entirely explained by these coarse length measures. TF-IDF remains at 0.981, so matching does not eliminate lexical cues. The caliper substantially reduces sample size and changes the evaluation population; these figures should be read as a targeted diagnostic, not as an estimate for all BonaFide traces. Matching also cannot control unobserved differences in hint content or writing style.

Figure 3 places these numbers in context. The step-count panel shows the large class difference that a model could exploit. The occupancy panel plots all held-out traces and shows substantial class overlap, avoiding the impression that a geometric quadrant separates the labels by itself. The per-model panel shows that SemFaith's AUROC ranges from 0.780 (Olmo-3-7B) to 0.977 (Qwen3-4B); the range cautions against treating the five-model mean as a universal constant. The supplementary appendix gives exact held-out counts and scores.

4.5 Feature families and hint constructions

The $K = 5$ feature-family ablation in Table 2 probes which measurements support the geometric result under the same five folds. Removing occupancy features lowers AUROC from 0.859 to 0.757, the largest change. Removing outlier features gives 0.826; removing adjacent-step coherence gives 0.856. Occupancy is therefore important to this predictor, while coherence contributes little

Table 2: Feature-family ablation at $K = 5$, using the five strict held-out-model folds. Δ is relative to all eight features.

Feature set	AUROC	Δ
All eight features	0.859	—
Without outlier distance	0.826	-0.033
Without occupancy	0.757	-0.102
Without coherence	0.856	-0.003

Table 3: AUROC within step-count strata, using unchanged held-out-model scores. Strata contain different class proportions; the long stratum has only six positive traces.

Steps	N	Pos.	Length	SemFaith	TF-IDF
≤ 5	161	84	0.642	0.653	0.993
6–10	211	78	0.781	0.802	0.987
≥ 11	290	6	0.989	0.962	0.850

to AUROC in this configuration. The features can nevertheless correlate, so an ablation measures predictive dependence of the fitted model rather than a unique mechanistic cause.

We also inspect AUROC within the seven recorded hint categories using the unchanged held-out-model predictions (see the supplementary appendix). The estimates span 0.758 for *validator* to 0.987 for *error_message*. Their positive counts differ sharply: *validator* has five acknowledged traces and *error_message* has three. Thus the table indicates that the overall score is not confined to one category, but it cannot establish transfer to an unseen hint mechanism. That would require holding out the hint type during training and ensuring both classes remain adequately represented in each test set.

4.6 Where the geometric signal comes from

The geometry is informative, but the intuitive direction of individual occupancy measurements requires care. Across the full evaluated set, acknowledged CoTs average 0.241 normalized $K = 5$ occupancy entropy, compared with 0.321 for concealed CoTs; their mean maximum-cluster shares are 0.798 and 0.779 respectively. The acknowledged class is shorter and thus has fewer opportunities to occupy different clusters. These aggregate directions do not support the simple narrative that acknowledgment always yields more dispersed trajectories. Figure 2 illustrates selected possibilities, while Figure 3 shows the overlap and length dependence of the complete sample. SemFaith's result comes from the fitted combination of distances, continuity, and occupancy at three resolutions, not from a universal high-entropy rule.

Table 3 further examines saved held-out predictions within step-count ranges. For traces of at most five steps, geometry reaches 0.653 AUROC, close to the 0.642 length control; the 6–10-step range yields 0.802 versus 0.781. Among traces with at least 11 steps, only six acknowledged examples remain, and the high AUROCs are unstable indicators of performance in that region. The TF-IDF predictor stays strongest in the two better-populated ranges. This stratification complements the one-to-one length match: both suggest a modest geometric signal in ranges with useful class overlap, while revealing how strongly the overall benchmark couples label and trace length.

Taken together, the feature ablation, matched pairs, and length strata locate the contribution more precisely. SemFaith offers a structured representation that generalizes across the five Instruct models and has residual predictive value after controlling major scale differences. The data do not warrant treating any single feature as a direct marker of concealed reasoning, and the remaining lexical shortcut invites new benchmarks with closer control of acknowledgment wording.

4.7 What multiple resolutions contribute

The three resolutions measure the same step sequence against increasingly fine training-derived partitions. Their joint score of 0.885 exceeds $K = 5$ (0.859), $K = 10$ (0.842), and $K = 20$ (0.854) under identical held-out folds. This gain does not come from a larger input trace or access to the audited model; it comes from retaining several views of where steps lie relative to shared reference spaces. Occupancy at $K = 5$ carries substantial predictive information, as the ablation shows, while finer partitions can distinguish traces that look similar at a coarse scale. The 24-dimensional description remains compact enough to inspect and compare with explicit alternatives.

The controlled evaluation sharpens that result. Geometry retains rank information among close length matches, and adding length to the geometric vector further increases full-fold AUROC to 0.912. These results make multi-resolution step geometry an empirically useful representation of BonaFide traces. TF-IDF remains stronger on this dataset; its near-ceiling result directs the next question toward settings where lexical acknowledgment cues are less available. The supplementary appendix records the precise subset and fitting rules, together with the earlier analyses excluded from the present evidence.

5 Discussion and Conclusion

SemFaith turns a CoT into a small collection of measurements of where its steps lie and how they move in a fixed embedding space. On BonaFide's hint-acknowledgment task, its 24-dimensional representation predicts labels across held-out Instruct models at 0.885 mean AUROC. The length-restricted analysis suggests that the score contains information beyond gross verbosity, while the multi-resolution comparison shows a gain over each tested single clustering resolution. These findings make step geometry a useful, inspectable axis for analyzing reasoning text.

The same experiments set a clear limit on the interpretation. BonaFide's label is tied to whether a misleading hint is acknowledged, and a TF-IDF model predicts it exceptionally well. Neither a geometric score nor an accurate lexical score can show which hidden computation produced an answer. Our visualizations likewise show textual patterns, not causal pathways. The current evidence therefore supports SemFaith as a predictor of this observable failure mode. Testing whether its features transfer to other definitions of faithfulness, including counterfactual or intervention-based evaluations, requires aligned labels on those tasks.

The measured dependence on step count and cluster occupancy points to concrete next tests. One can construct paired traces with similar wording and length but different causal histories, hold out hint mechanisms during training, and compare geometry with lexical and pooled-embedding controls on the same examples. Such experiments would identify when a structural representation adds information that surface text does not already supply. The present contribution is the representation, a corrected model-held-out assessment, and explicit controls that keep the claim proportional to the evidence.

Reproducibility Statement

The supplementary code contains step segmentation, fixed sentence encoding, fold-wise clustering and scoring, length and lexical controls, matching analysis, plotting scripts, and per-fold and per-example outputs. The public BonaFide dataset and encoder are identified in the code README. The supplementary appendix reports exact model counts, feature definitions and protocol details. The original main experiment can be rerun separately; it is not combined with the strict five-model comparison.

AI Use Statement

Generative AI tools assisted with language editing, presentation. During revision, they also assisted with the design and implementation of additional evaluation controls and interpretation of their outputs. The authors verified the code, numerical results, figure content, and final claims and take responsibility for the submission.

Ethics Statement

This study analyzes model-generated reasoning traces from a public benchmark. It collects no new human-subject data and does not require access to the audited models' internal states. Because a text-based score may be mistaken for a causal measure of reasoning, we state its operational target and limitations explicitly.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008.

Fabian Fumagalli, R Teal Witter, and Christopher Musco. Polyshap: Extending kernelshap with interaction-informed polynomial regression. *arXiv preprint arXiv:2601.18608*, 2026.

Yoav Gur-Arieh, Ana Marasović, and Mor Geva. Faithfulness metrics don't measure faithfulness: A meta-evaluation with ground truth. *arXiv preprint arXiv:2605.25052*, 2026.

Yunseok Han, Yejoon Lee, and Jaeyoung Do. Rfeval: Benchmarking reasoning faithfulness under counterfactual reasoning intervention in large reasoning models. *arXiv preprint arXiv:2602.17053*, 2026.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4198–4205, 2020.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024.

Katie Matton, Robert Ness, John Guttag, and Emre Kiciman. Walk the talk? measuring the faithfulness of large language model explanations. In *International Conference on Learning Representations*, volume 2025, pp. 73212–73277, 2025.

Avni Mittal and Rauno Arike. C2-faith: Benchmarking llm judges for causal and coverage faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2603.05167*, 2026.

Filip Naudot, Tobias Sundqvist, and Timotheus Kampik. llmshap: A principled approach to llm explainability. *arXiv preprint arXiv:2511.01311*, 2025.

Beatrix Miranda Ginn Nielsen and Lars Kai Hansen. Hubness reduction improves sentence-bert semantic spaces. In *Northern Lights Deep Learning Conference*, pp. 181–204. PMLR, 2024.

Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6048–6089, 2024.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15012–15032, 2024.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 410–420, 2007.

David Sculley. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pp. 1177–1178, 2010.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024.

Xu Shen, Song Wang, Zhen Tan, Laura Yao, Xinyu Zhao, Kaidi Xu, Xin Wang, and Tianlong Chen. Faithcot-bench: Benchmarking instance-level faithfulness of chain-of-thought reasoning. *arXiv preprint arXiv:2510.04040*, 2025.

Matthieu Tehenan. Semantic geometry of sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11993–12004, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Chase Walker and Rickard Ewetz. Explaining the reasoning of large language models using attribution graphs. *arXiv preprint arXiv:2512.15663*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Kerem Zaman and Shashank Srivastava. A causal lens for evaluating faithfulness metrics. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29413–29437, 2025.