The Geometry of Reasoning Failure: Predicting Agent Errors from Semantic Scale Trajectories

Nikita Kazeev [*1] **Andrey Ustyuzhanin** [*2 3 1]

Abstract

Selective prediction over agentic LLM outputs needs uncertainty signals. Direct examination of Chain-of-thought is a natural approach, but the reasoning text often becomes misleading and unreliable on the difficult tasks, where it's needed most. We introduce Semantic Level of Detail (SLoD), a continuous macro-to-micro abstraction axis over reasoning chunks, recoverable through three independent operationalizations (a SciBERT linear probe, a probe-free embedding-axis projection, and pairwise LLM judging), and use the shape of an agent's SLoD trajectory as a black-box selective-prediction signal. A LightGBM classifier over a stack of interpretable trajectory detectors predicts SWE-agent failures at ROC-AUC 0.89 ± 0.02 and, used as an out-of-fold scorer to abstain from low-quality attempts, matches a plurality self-consistency vote and an LLM self-verification baseline on the Frontier-Science Olympiad subset – but from a black-box, answer-agnostic signal – and its coverage–accuracy curve dominates a length-only baseline at every operating point. At the natural best-of-5 budget it lifts Pass@1 from 58.2% to 68.0% over five DeepSeek-V3.2 attempts. The same trajectory features classify hallucination categories on the OpenManus subset of AgentHallu at macro-AUC 0.71. We also report negative results on GPQA-Diamond and Qwen3-30B traces, showing that the signal is conditional on the generator and the problem.

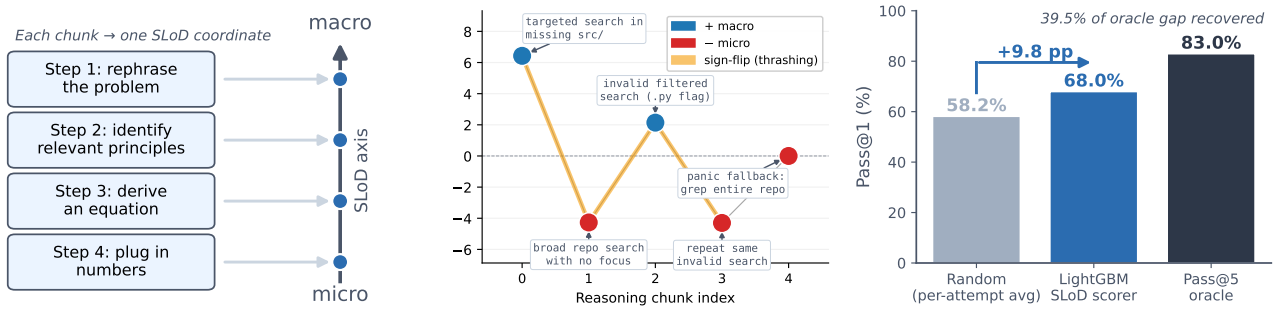
[1]Institute for Functional Intelligent Materials, National University of Singapore, Block S9, Level 9, 4 Science Drive 2, Singapore 117544 [2]Constructor University, Bremen, Campus Ring 1, 28759, Germany [3]Constructor Knowledge Labs, Bremen, Campus Ring 1, 28759, Germany. Correspondence to: Nikita Kazeev <kazeevn@gmail.com>, Andrey Ustyuzhanin <Andrey.Ustyuzhanin@constructor.org>.

Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

Large language models deployed as autonomous agents – writing code, solving scientific problems, and orchestrating tools – fail in ways that are difficult to detect from the final output alone. Recent analyses of multi-agent systems catalog more than a dozen distinct failure modes spanning system design, inter-agent miscoordination, and verification breakdowns (Cemri et al., 2025), and chain-of-thought (CoT) reasoning, despite its empirical gains, generalizes poorly out of distribution and depends sensitively on prompt engineering rather than on transferable algorithmic procedures (Stechly et al., 2024). The natural surface for diagnosis – the model's own intermediate tokens – has proved unreliable: token-level hallucination cues are obscured precisely when CoT is applied, and the semantic content of CoT traces is only loosely coupled to the underlying computation. This motivates a different framing: rather than reading what the model *says* during reasoning, can we measure how the reasoning *moves*, and use that movement as an uncertainty signal we can abstain on?

We pursue this question through a single geometric construct, Semantic Level of Detail (SLoD): a continuous abstraction axis from broad/contextual (*macro*) to specific/numerical (*micro*) along which any reasoning chunk can be located. Applying it to the chunks of an agent's reasoning trace yields a one-dimensional time series – a "line of thought" through abstraction space (Sarfati et al., 2025) – whose shape, we argue, carries practical information about whether the agent will succeed. SLoD is black-box (a frozen sentence encoder or a non-reasoning LLM judge suffices) and trace-aware, so it applies uniformly to closed models and complements white-box logit-based uncertainty signals. Figure 1 provides the general overview: by projecting reasoning chunks onto the abstraction axis we obtain a trajectory whose shape both discriminates successful from failed agent traces and supports selective sampling at fixed budget that reaches parity with self-consistency and self-verification baselines from a black-box, answer-agnostic signal. Figure 2 shows that the signal lives in the *shape* of the trajectory, not its average level; Section 3.4.1 further shows that on SWE-agent it is not reducible to trace length or truncation either.



(a) Reasoning chunks map to a one-dimensional SLoD trajectory.

(b) A failed SWE-agent trajectory on `ReproNim/reproman-518`.

(c) Trajectory-based reranking lifts FrontierScience Pass@1.

Figure 1. Overview of the approach. Panel (a) shows how reasoning chunks map to a one-dimensional SLoD trajectory. Panel (b) shows a concrete SWE-agent failure (`ReproNim/reproman-518`, llama-70B, *exit_context*) repeatedly invokes a broken `search_dir` on a non-existent `src/` directory, oscillating between high-level framing (blue, +macro) and detail-level retries (red, −micro) on 4 of 4 chunk boundaries; the `thrashing` and `no_commitment` detectors fire on exactly this signature. Panel (c) shows that on FrontierScience Olympiad with five independent DeepSeek-V3.2 attempts, an out-of-fold LightGBM scorer over trajectory features lifts Pass@1 from 58.2% to 68.0%, on par with the self-consistency and self-verification baselines (Table 4).

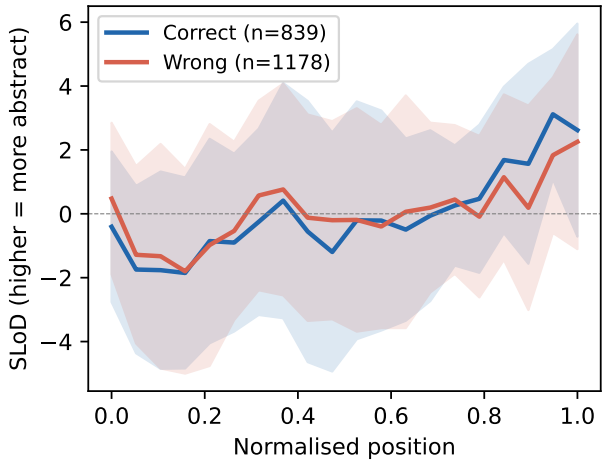


Figure 2. Mean SLoD trajectory over SWE-agent reasoning traces, separated by outcome. Higher values indicate more abstract chunks; position 0–1 is the normalized location within the trace; shaded bands are $\pm 1\sigma$. The average trajectories are indistinguishable, yet as we show in Section 3.4.1 the shape of the trajectory carries a strong failure signal.

1.1. Related Work

Our work is positioned between two existing lines. Recent white-box methods predict CoT correctness from structural fingerprints in a model's internal computational graph (Zhao et al., 2026) and require per-model transcoders; test-time selection methods rerank multiple samples using cheap output-side signals such as token self-certainty for closed models (Kang et al., 2026) but ignore the trace altogether.

LLMs provide a natural ground for estimating uncertainty – token probabilities. Zhu et al. (2026); Zhao et al. (2025); Zhang et al. (2026) use token-probability signals along the reasoning trajectory to estimate error likelihood: Zhu et al. (2026) tie token-entropy spikes to failure onsets, Zhao et al. (2025) propagate per-step entropy with situational weights, and Zhang et al. (2026) mine interpretable trajectory-shape features from the token-log-probability series and fit a simple calibrator – the closest methodological neighbour to our approach. These methods are powerful but bounded by their signal source: being downstream of token probabilities, they require access to per-token log-probabilities (full logits, or API-exposed log-probs) and are, in principle, blind to "unknown unknowns" that are confidently wrong at the token level. SLoD trajectories instead read the *abstraction shape* of the trace: they need only a frozen sentence encoder or a non-reasoning LLM judge, apply uniformly to closed models that we cannot introspect, and are answer-agnostic, so they complement rather than replace log-probability methods such as the trajectory-feature calibrator of Zhang et al. (2026).

What is new here. Reading an agent's reasoning trajectory as an uncertainty signal is not itself new; our contribution is a distinction in *signal source*. Every method above derives its per-step quantity from the generator's own token distribution and asks *how confident* the model is. SLoD instead measures the *semantic content* of each chunk – its location on an externally defined macro-to-micro abstraction axis – and asks *at what level of abstraction* the reasoning operates. This is a different latent property, not a cheaper estimator of the same one, with three consequences: *(i)* the signal is orthogonal by construction to token-probability uncertainty, so it can in principle flag confidently-wrong failures that low token-entropy conceals; *(ii)* it needs no access to the generator's token distribution – a frozen encoder or a non-reasoning judge over text suffices – so it is

defined where logit-based methods are not: our strongest result is on logged third-party SWE-agent traces (Badertdinov et al., 2024), for which no token log-probabilities exist, and more generally reasoning models served through closed APIs typically do not expose log-probabilities for their chain-of-thought tokens – the very series a calibrator like Zhang et al. (2026) requires; on such surfaces SLoD remains computable (SWE-agent AUC 0.893), whereas a logit-trajectory method cannot be run at all; and *(iii)* because the signal is external and content-based rather than tied to one model's logits, the same construct admits three independent operationalizations (probe, axis, judge) that converge (Section 2.2.1) – evidence that SLoD is a property of reasoning text rather than an artifact of a single estimator. We therefore frame SLoD not as a competitor to token-probability calibration but as a complementary axis: combinable where both are available, and still defined where token probabilities are not.

The broader uncertainty-quantification literature for LLMs offers several reference points. Semantic entropy clusters paraphrastic samples to estimate generation uncertainty without ground truth (Kuhn et al., 2023); self-consistency selects the modal answer across independent samples and is the canonical sampling-based baseline (Wang et al., 2023); conformal prediction provides distribution-free coverage guarantees for LLM outputs under exchangeability (Quach et al., 2024). Process reward models (Lightman et al., 2024) score reasoning *step correctness* from supervised step-level labels and require a learned per-step verifier; SLoD trajectories instead score reasoning *shape* from a single scalar per chunk and require no step-level supervision. We view SLoD as a complementary black-box selective-prediction signal that can be combined with – not replaced by – any of these methods.

The results presented in this paper are also discussed in a non-archival workshop paper (Kazeev & Ustyuzhanin, 2026).

2. Semantic Level of Detail (SLoD)

We define Semantic Level of Detail (SLoD) as a continuous abstraction axis from broad/contextual (*macro*) to specific/numerical (*micro*). We formalize and operationalize SLoD in three ways: (i) a linear probe over frozen SciBERT (Beltagy et al., 2019) `[CLS]` embeddings, trained on heuristic section-level labels from scientific text (Section 2.1); (ii) a probe-free axis projection in the same embedding space, defined as $\mu_{\text{micro}} - \mu_{\text{macro}}$ in the spirit of difference-in-means probes that have proved effective for truth-direction discovery (Marks & Tegmark, 2023) (Section 2.2); and (iii) pairwise LLM judging with Plackett–Luce (Plackett, 1975; Luce et al., 1959) aggregation, using no embeddings (Section 2.3). The agent-failure experiments in Section 3 use the pairwise variant, which is robust to domain shift and is the operationalization on which all reported AUCs are computed. We include the embedding variants because they are cheaper, more mechanistic, and provide independent corroboration that the SLoD axis is a real property of reasoning text rather than an artifact of any one judge; Section 2.2.1 reports an embedding-axis experiment on chain-of-thought traces that validates this claim independently.

2.1. SciBERT linear probe

We embed scientific spans with frozen SciBERT and train a multinomial logistic regression on the `[CLS]` vector to predict the heuristic section-level macro/meso/micro label assigned from QASPER document structure (section title and depth). The probe is trained with L2 regularization on a length-matched, balanced split of $37\,278$ spans. On a held-out test split this configuration achieves macro-F1 $= 0.720$ (random baseline 0.328, word-count-only baseline 0.263), confirming that abstraction level is linearly decodable from frozen embeddings rather than being a length artifact. Per-class F1 scores are macro 0.821, meso 0.616, micro 0.724 – the meso class is the harder middle band, as expected.

2.2. Probe-free axis projection

The probe is convenient but discretizes the signal and inherits the errors of the meso class. We therefore also construct a probe-free axis: in the same SciBERT embedding space, define

$$\mathbf{v}_{\text{SLoD}} = \frac{\mu_{\text{micro}} - \mu_{\text{macro}}}{\|\mu_{\text{micro}} - \mu_{\text{macro}}\|},$$

where μ_c is the mean embedding of training spans labeled c. The per-chunk SLoD coordinate is the inner product of its `[CLS]` embedding with $\mathbf{v}_{\text{SLoD}}$ – one forward pass through frozen SciBERT plus a linear operation, with no learned parameters at inference. This is the construction of Marks & Tegmark (2023) for truth direction transferred to the abstraction axis.

2.2.1. Validation on QASPER chain-of-thought

Before applying SLoD to agent traces, we validate the embedding-based axis on $6{,}101$ chain-of-thought reasoning steps from $2{,}000$ Claude-Haiku traces over QASPER. On a held-out test split of the heuristically labeled spans the centroid axis cleanly orders the three levels (Cohen's $d = 2.65$ between macro and micro; class means -0.534, 3.166, 4.710). On the CoT data the mean projection of a trace's steps onto $\mathbf{v}_{\text{SLoD}}$ correlates with answer attribution-F1 at Spearman $\rho = +0.219$ (Bonferroni-corrected $p < 10^{-21}$), beating the strongest discrete-probe baseline ($\rho = -0.197$ for transition-matrix features). A logistic regression over seven SLoD-axis features reaches AUC 0.615 on attribution-

F1, while a control regression over three orthogonal embedding directions reaches only 0.549 – confirming the signal is specific to the macro–micro direction, not to generic embedding variance. The strongest predictor is the static mean projection, not its dynamics, suggesting that on QASPER *the level at which a trace operates* matters more than *how it moves*.

2.3. Pairwise LLM judging

For agent traces we drop the embedding pipeline entirely. Each trace is split into reasoning and answer chunks; chunks within a single trace are ranked by repeated pairwise LLM comparison ("Which chunk is at a more specific/concrete level of detail?"), and the per-chunk SLoD coordinate is the Plackett–Luce rating fit from the wins, losses, and ties. This variant requires only a non-reasoning judge LLM and produces a single scalar SLoD coordinate per chunk, normalized within each trace. Pair selection prioritizes informative comparisons. DeepSeek-V3.2 serves as the SLoD judge for all DeepSeek-V3.2 and SWE-agent runs; for Qwen3-30B traces, Qwen3-30B-A3B-Instruct-2507 is used as its own judge to keep the SLoD pipeline fully local.

2.4. Trajectory features

For each trace we obtain two ordered SLoD series: the reasoning trajectory (over reasoning chunks) and, where the dataset distinguishes them, the answer trajectory (over answer chunks). From each series we derive a fixed set of summary features capturing length, level, shape, monotonicity, and thrashing – e.g. `n_chunks`, `start`, `end`, `mean`, `range`, `monotonicity`, `total_variation`, `direction_changes`, plus a 20-point time-resampled series `traj_t00`–`traj_t19` for shape – giving 47 trajectory features for SWE-agent (reasoning only) and up to 99 for datasets with both reasoning and answer trajectories (degenerate features with no within-fold variance are dropped per dataset, so the realised count can be lower on small slices).

2.5. Failure prediction protocol

We predict a binary outcome: whether the model successfully solved the task, according to the ground-truth labels in the dataset. We use stratified 5-fold cross-validation. Each fold is fit independently on three nested feature sets:

`length_only` / `length_abort`. A 1–3 feature structural baseline: chunk counts of the reasoning, answer, and full trace. On SWE-agent we also add a binary variable indicating whether the agent execution was terminated prematurely.

`trajectory_full`. All 47/99 trajectory features, including the chunk-count features.

`trajectory_shape`. The same trajectory features with the chunk-count features removed, isolating trajectory shape from trace length.

`mode_stack`. A 4–13-dimensional vector of interpretable detector scores. Core detectors include *thrashing*, *no commitment*, *rambling/overlong*, *premature exit*, *late derailment*, and *truncation abort*, plus answer-side variants where applicable. Dimensionality varies by dataset: 4 when only an answer trajectory exists (Qwen runs with no reasoning chunks), 7 on SWE-agent (reasoning detectors plus `truncation_abort`), and 13 when reasoning and answer trajectories are both available. Each detector is a fixed function of trajectory features with no learned parameters; the `mode_stack` vector handed to the classifier is these raw per-trace detector scores, computed independently for each trace without reference to labels or to other traces, so every reported AUC is strictly out-of-fold. Each detector additionally encodes a directional hypothesis (higher score $\Rightarrow$ more likely to fail) which we test with a 95% percentile-bootstrap CI on its single-feature failure AUC and label as *confirmed* (CI lower bound > 0.5), *inverted* (CI upper bound < 0.5), *inconclusive* (CI straddles 0.5), or *insufficient_data*; this falsifiability check thresholds each score against the 90th percentile of the success-trajectory distribution, a calibration that enters only the verdict table and never the predictive features.

We fit each feature set with two classifiers. The **logistic regression** variant is a median-imputed, standardised pipeline with `max_iter=5000` and balanced class weights. The **LightGBM** variant uses median imputation and `LGBMClassifier` (300 trees, 31 leaves, learning rate 0.05, `min_child_samples=10`, balanced class weights). The same protocol is shared across every dataset and run reported in Section 3; all feature engineering, detector calibration, and CV splitting happen inside the SH6 failure-analysis module.

3. Experiments

3.1. Models

We evaluate two open-source models: DeepSeek-V3.2 (Liu et al., 2025) (accessed through API) and Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025) (ran locally). DeepSeek-V3.2 is the SLoD pairwise judge for DeepSeek-V3.2 and SWE-agent runs; Qwen3-30B-A3B-Instruct-2507 is its own judge for Qwen runs (Section 2.3).

3.2. Datasets

FrontierScience (Wang et al., 2026). The 160-example open-source slice contains two tracks: *Olympiad* (international olympiad-level problems: IPhO, IChO, IBO; 100 items) and *Research* (PhD-level open-ended problems; 60

items). For each model we generate traces, follow the FrontierScience LLM-as-a-judge grading protocol against the reference answer, and compute SLoD trajectories with the pairwise judge. *All FrontierScience numbers in this paper are reported on the Olympiad subset only.* DeepSeek-V3.2 solves $0/60$ Research problems across five seeds, so the aggregate FrontierScience accuracy is dominated by Olympiad and reporting it would obscure the actual difficulty distribution. Per-attempt valid-trace counts average $96/100$ (DeepSeek) and $98/100$ (Qwen3-30B) of the 100 Olympiad items after dropping attempts that produced no final answer.

SWE-agent trajectories (Badertdinov et al., 2024). Software engineering agent trajectories from the SWE-agent framework. We sample 2017 trajectories total split across three generators: llama-8b (n=549), llama-70b (n=1,232), and llama-405b (n=236); the dataset ships success labels and we use DeepSeek-V3.2 to compute SLoD. SWE-agent trajectories have no explicit answer chunks, so only reasoning-side features and detectors apply.

GPQA-Diamond (Rein et al., 2023). Graduate-level multiple-choice questions in biology, physics, and chemistry. The Diamond subset is the 198-item slice on which both domain experts agree on the correct answer but a majority of non-experts fail. For each model we generate traces, compare the model's selection to the reference answer, and compute the SLoD trajectory.

AgentHallu (Liu et al., 2026) (OpenManus subset). We use the OpenManus framework slice of AgentHallu, which contains agent traces labeled with hallucination categories. OpenManus is the only AgentHallu framework slice with enough labeled hallucination-category failures to support a 3-class classification problem at non-degenerate per-class counts; the other framework slices each contain too few items per category for a stratified 5-fold protocol. Following the framework's released labels, we treat the 84 failure items as a 3-class classification problem (*Human-Interaction*, *Reasoning*, *Retrieval* hallucination), and additionally evaluate binary failure prediction on the full 104-item slice.

3.3. Metrics

The primary evaluation is binary failure prediction: ROC-AUC under stratified 5-fold CV. For the OpenManus hallucination-category task we report macro one-vs-rest AUC. For FrontierScience Olympiad we additionally evaluate inference-time selection: each problem is solved $5\times$ with DeepSeek-V3.2; we report Pass@1 (mean across attempts), Pass@5 (oracle: any attempt correct), and SLoD-selected Pass@1 using strictly Olympiad-only out-of-fold scorer predictions.

Model	ROC-AUC
`length_abort` (logreg)	0.68 ± 0.02
`trajectory_shape` (logreg)	0.82 ± 0.02
`trajectory_full` (logreg)	0.82 ± 0.02
`trajectory_full` (lightgbm)	0.82 ± 0.02
`mode_stack` (logreg)	0.70 ± 0.02
`mode_stack` (lightgbm)	$\mathbf{0.89 \pm 0.02}$

Table 1. SWE-agent aggregate failure prediction, n=2017 trajectories (839 correct / 1178 wrong), stratified 5-fold CV. `trajectory_shape` drops the chunk-count features, isolating shape from length: it matches `trajectory_full` and still beats the `length_abort` baseline – which already bundles chunk counts and the truncation flag – by $+0.14$.

3.4. Results

We organise results by surface. Section 3.4.1 (SWE-agent) and Section 3.4.2 (FrontierScience Olympiad) report binary failure-prediction AUC; Section 3.4.3 casts the same scorer as a selective-sampling signal and is the most directly operational result for selective-prediction settings; Section 3.4.4 extends to hallucination-category classification on AgentHallu/OpenManus; Section 3.4.5 reports a clean negative on GPQA-Diamond.

3.4.1. SWE-AGENT FAILURE PREDICTION

Table 1 reports cross-validated AUCs over the full pool of 2,017 traces. Trajectory features beat the structural baseline by a wide margin, and the lift is not a length or truncation artifact: the `length_abort` baseline already bundles both the chunk-count features and the truncation flag, yet `trajectory_full` logreg improves on it by $\Delta\text{AUC} = +0.147$ (0.822 vs. 0.675), and a chunk-count-free shape-only feature set reaches the same 0.822 – so this gain is carried by trajectory shape, not by trace length. The strongest single classifier on this dataset is LightGBM over the 7-dimensional interpretable detector stack (`mode_stack`), reaching AUC 0.893 ± 0.018 – the headline SWE-agent number. That stack additionally includes `truncation_abort`, the single strongest detector here (failure AUC 0.648), so the increment from the 0.822 shape-only model to 0.893 partly reflects truncation; the $+0.147$ shape lift over `length_abort` is the length- and truncation-independent core of the result.

The aggregate hides a strong per-generator effect (Table 2). On llama-70b traces both `trajectory_full` (0.896) and `mode_stack` LightGBM (0.907) decisively beat the structural baseline (0.632). On llama-8b and llama-405b the picture is different: the `length_abort` baseline is already strong (0.759 and 0.753 respectively), `trajectory_full` slightly underperforms it, and only the LightGBM `mode_stack` reliably surpasses it (0.850

and 0.827). Two interpretations are consistent with this pattern: (i) for very small or very large generators a non-trivial fraction of failures are explained by trace-truncation length signals that the trajectory shape cannot add to; (ii) the larger middle-tier llama-70b traces fail in SLoD-shape-discriminable ways at higher rate.

3.4.2. FrontierScience Olympiad failure prediction

Per-attempt failure prediction on the Olympiad subset is reported in Table 3 for both models. On DeepSeek-V3.2 (n=95, 61/34) the LightGBM `trajectory_full` classifier reaches AUC 0.699, beating the length baseline 0.663, and the logreg `mode_stack` reaches 0.733. For Qwen3-30B (n=98, 49/49), trajectory features collapse to chance (0.526 logreg, 0.509 LightGBM); only the length baseline (0.708) and `mode_stack` (0.700) beat chance, indicating that the SLoD signal does not transfer to Qwen3-30B traces under this protocol.

3.4.3. FrontierScience Olympiad: selective sampling under fixed budget

This is the selective-prediction view of SLoD: a per-attempt scorer used to abstain from low-quality samples in a fixed-budget sampling regime. We lead with the two operating-point-independent facts rather than any single number: (i) the full coverage–accuracy curve (Figure 3) dominates a length-only baseline at *every* coverage, and (ii) at the natural best-of-5 budget the SLoD scorer reaches parity with the self-consistency and self-verification sampling baselines (Table 4). Although the Olympiad failure-prediction AUC of Table 3 is moderate (0.74), the same scorer drives these selective gains – a moderate-AUC scorer can still produce large selective gains when the score is monotone with attempt quality. We sample five DeepSeek-V3.2 attempts per Olympiad problem (`reasoning-auto` with seeds $s_0 \dots s_4$), score each attempt with a `trajectory_full` classifier trained on the Olympiad subset under strict out-of-fold scoring, and select the highest-scoring attempt per problem. Unscored attempts rank last during selection. Inference/grading errors count as failed attempts. Table 4 reports the headline at coverage $1/5$ (one selected attempt per problem): the LightGBM scorer raises Pass@1 from the 58.2% per-attempt average to 68.0%, a $+9.8$ pp gain that recovers 39.5% of the gap to the 83.0% Pass@5 oracle. Logistic regression on the same features lifts only by $+1.8$ pp, indicating that the selective signal lives in non-linear feature interactions.

Sampling-based baselines. Two canonical test-time-selection methods choose among the *same* five attempts and are therefore directly comparable (Table 4). *Self-consistency* (Wang et al., 2023) takes the plurality vote over the five normalized answers (ties broken by run order, so the rule is label-free); *LLM self-verification* (Weng et al., 2023) re-grades each attempt with the same DeepSeek-V3.2 model – shown the problem, reasoning, and final answer but *not* the reference – and retains the attempt with the highest signed confidence (the verifier's own correct/incorrect call agrees with the ground-truth grade 72–79% of the time, OOF AUC 0.73–0.80). Self-consistency reaches 68.0% Pass@1 and self-verification 67.0%; the SLoD LightGBM scorer (68.0%) matches the former and edges the latter. On this closed-answer Olympiad benchmark SLoD selective sampling is thus *competitive with, but does not beat,* self-consistency. The methods are complementary rather than redundant: self-consistency requires comparable answers and is undefined on the open-ended Research track (where DeepSeek-V3.2 scores $0/60$ and no plurality forms), and even on Olympiad its winning answer captures only $2.4/5$ votes on average ($1.6/5$ on physics); SLoD instead scores each attempt's trajectory shape in isolation and needs no inter-sample agreement, while self-verification needs an extra grading call per attempt. We read the headline not as SLoD dominating sampling-based selection but as a black-box, answer-agnostic signal reaching parity with it.

Figure 3 extends Table 4 from a single operating point to the full coverage–accuracy curve. Pooling the ~ 480 DeepSeek-V3.2 OOF attempt scores across the five seeds, we sort attempts by SLoD-LightGBM score and retain the top-c fraction. The SLoD curve dominates a length-only logistic-regression baseline (3 chunk-count features, same 5-fold OOF protocol) at *every* coverage operating point: at $c \approx 0.4$ the SLoD selector reaches 74.6% Pass@1 vs. 67.2% for the length baseline and 60.2% for uniform random retention.

3.4.4. AgentHallu/OpenManus hallucination categories

On the 84-item OpenManus failure subset of AgentHallu we cast hallucination categorisation as a 3-class classification problem (n_{HIH}=42, $n_{\text{Reasoning}}$=20, $n_{\text{Retrieval}}$=22). Stratified 5-fold CV macro-AUC under `trajectory_full` reaches 0.712 with LightGBM and 0.682 with logistic regression, against 0.520 for `length_reasoning_only`. Adding answer-chunk count raises the `length_only` baseline to 0.671. Per-class LightGBM AUCs are 0.790 for Retrieval, 0.771 for Reasoning, and 0.575 for Human-Interaction. Paired-bootstrap analysis shows that the lift over `length_reasoning_only` is statistically significant ($\Delta = +0.194$, CI $[+0.081, +0.299]$) but the lift over `length_only` is *not* ($\Delta = +0.041$, CI $[-0.055, +0.142]$, straddles 0): once the answer-chunk count is in the baseline, trajectory shape adds no statistically detectable discrimination. The OpenManus result is therefore scoped to *which* hallucination occurred among already-failed traces, not to predicting failure itself: binary failure prediction on the

Model	8b	70b	405b
n	549	1232	236
`length_abort` (logreg)	0.76 ± 0.04	0.63 ± 0.03	0.75 ± 0.10
`trajectory_full` (logreg)	0.72 ± 0.05	0.90 ± 0.01	0.75 ± 0.06
`trajectory_full` (lightgbm)	0.71 ± 0.05	0.90 ± 0.01	0.76 ± 0.05
`mode_stack` (logreg)	0.75 ± 0.04	0.66 ± 0.03	0.74 ± 0.04
`mode_stack` (lightgbm)	$\mathbf{0.85 \pm 0.03}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.83 \pm 0.03}$

Table 2. SWE-agent per-generator ROC-AUC. Trajectory shape dominates on llama-70b but ties or loses to length-only on the extremes (llama-8b, llama-405b); only the LightGBM mode stack is a consistent winner across all three.

Model	DeepSeek-V3.2	Qwen3-30B
`length_only` (logreg)	0.66 ± 0.20	0.71 ± 0.06
`trajectory_full` (logreg)	0.61 ± 0.12	0.53 ± 0.08
`trajectory_full` (lightgbm)	0.70 ± 0.07	0.51 ± 0.14
`mode_stack` (logreg)	0.73 ± 0.13	0.70 ± 0.11
`mode_stack` (lightgbm)	0.74 ± 0.08	0.57 ± 0.08

Table 3. FrontierScience Olympiad failure prediction ROC-AUC.

Selector	Pass@1
Per-attempt average	58.2%
SLoD logistic-regression scorer	60.0%
LLM self-verification	67.0%
Self-consistency (plurality vote)	**68.0%**
SLoD LightGBM scorer	**68.0%**
Pass@5 oracle	83.0%

Table 4. FrontierScience Olympiad selective sampling over 5 DeepSeek-V3.2 attempts. All selectors retain one of the same five attempts per problem (coverage $= 1/5$). The SLoD LightGBM scorer ties the plurality self-consistency vote at 68.0% and edges LLM self-verification (67.0%); both recover $\approx$40% of the average-to-oracle gap.

full 104-item slice is dominated by length (`length_only` 0.638 vs. `trajectory_full` logreg 0.472). SLoD adds value in the *category* problem, not the *occurrence* problem on this slice.

3.4.5. GPQA-Diamond: a negative result

On GPQA-Diamond, trajectory features fail to beat the length baseline for both models (Table 5). For DeepSeek-V3.2 (n=192) `length_only` reaches AUC 0.761, while `trajectory_full` reaches only 0.706 (logreg) and 0.656 (LightGBM). For Qwen3-30B (n=197) the gap is even larger: 0.702 for length vs. 0.563 (logreg) and 0.620 (LightGBM) for trajectory. The `mode_stack` models (0.743 and 0.689 respectively) come close to the length baseline but never exceed it. We read this as a clean negative result: on GPQA-Diamond's multiple-choice format the SLoD trajectory shape carries essentially no information beyond reasoning length.

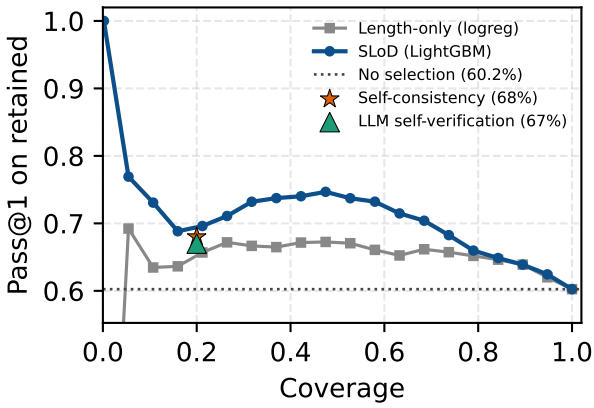


Figure 3. Coverage–accuracy curve for selective sampling on FrontierScience Olympiad ($\sim$ 480 DeepSeek-V3.2 attempts pooled across five seeds). At each coverage c, retain the top-c fraction of attempts ranked by score. SLoD-LightGBM (blue) dominates a length-only logreg (gray) at every operating point; the dotted line is uniform random retention (60.2%, the pooled per-attempt mean). The single-selection sampling baselines – self-consistency ($\star$, 68%) and LLM self-verification (▲, 67%) – each retain one attempt per problem and are therefore marked at the best-of-5 operating point (coverage $1/5$). OOF scores are pooled across five independent seed runs; each scorer is trained out-of-fold within its own seed.

4. Conclusion

We show that the *shape* of an agent's macro-to-micro reasoning trajectory – recovered without access to logits or activations – is an informative but conditional diagnostic for failure. Where the signal is present, it is large (SWE-agent AUC 0.893, FrontierScience Olympiad Pass@1 lift +9.8 pp, on par with a self-consistency vote); where it is not, length and answer-chunk count already saturate the predictable variance (GPQA-Diamond, Qwen3-30B, the OpenManus categorization lift over `length_only`).

Limitations *(i) In-dataset evaluation.* All failure-prediction AUCs are 5-fold CV within a single dataset, and the FrontierScience Olympiad rerank uses out-of-fold but in–dataset scoring; we do not yet validate transfer of a frozen scorer across datasets or generators. *(ii) Cross-model variance.* The Qwen3-30B traces collapse to chance under

Model	DeepSeek-V3.2	Qwen3-30B
n (+/−)	192 (147/45)	197 (127/70)
`length_only` (logreg)	$\mathbf{0.76 \pm 0.05}$	$\mathbf{0.70 \pm 0.08}$
`trajectory_full` (logreg)	0.71 ± 0.09	0.56 ± 0.05
`trajectory_full` (lightgbm)	0.66 ± 0.05	0.62 ± 0.05
`mode_stack` (logreg)	0.74 ± 0.06	0.69 ± 0.09
`mode_stack` (lightgbm)	0.70 ± 0.05	0.62 ± 0.06

Table 5. GPQA-Diamond ROC-AUC. Trajectory features fail to beat the length baseline (in bold).

trajectory features on FrontierScience Olympiad and underperform length on GPQA-Diamond, a clear negative result and a warning that SLoD trajectory shape is conditional on the generator's reasoning style.

Future work An activation-space probe would remove the LLM-judge dependency at inference time, enabling online steering rather than offline diagnosis. Computing SLoD on trace prefixes naturally suggests an anytime-valid formulation in which the score is treated as an evidence process and the agent abstains as soon as the score crosses a calibrated threshold; combined with the self-consistency and self-verification baselines of Section 3.4.3, which our scorer already matches offline, this is a clean route to online selective sampling over agentic LLMs. Two concrete next steps – cross-generator scorer transfer on SWE-agent and a prefix-length ablation – are scoped in detail in the follow-up notes accompanying this work. Finally, instrumenting more generators (beyond DeepSeek-V3.2 and Qwen3-30B) on text-answer benchmarks is needed to map the generator-conditioning of the trajectory-shape signal that Section 3.4.1 and Section 3.4.2 expose.

A direct comparison with logit-based uncertainty is a further priority. Methods that read per-token entropy or log-probabilities along the trajectory (Zhu et al., 2026; Zhao et al., 2025; Zhang et al., 2026) measure the generator's confidence, whereas SLoD measures semantic abstraction; the two are constructed from disjoint information, so we expect a stacked SLoD+logit scorer to be complementary rather than redundant. The comparison is only defined where full logits are available, which excludes our headline logged-trace and closed-API settings; the most accessible venue is our locally-hosted Qwen3-30B, which is also the regime where SLoD underperforms length (Section 3.4.2). Pairing the two there would test whether Qwen's failures are length-saturated for *any* signal or specific to abstraction shape, and whether the signals add when combined – turning the present conditional negative into a sharper map of when trajectory-shape information is worth extracting.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Badertdinov, I., Trofimova, M., Anapolskiy, Y., Abramov, S., Zainullina, K., Golubev, A., Polezhaev, S., Litvintseva, D., Karasik, S., Fisin, F., Skvortsov, S., Nekrashevich, M., Shevtsov, A., and Yangel, B. Scaling data collection for training software engineering agents. *Nebius blog*, 2024.

Beltagy, I., Lo, K., and Cohan, A. SciBERT: A pretrained language model for scientific text. 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. 3615–3620, 2019.

Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K., Zaharia, M., Gonzalez, J. E., and Stoica, I. Why Do Multi-Agent LLM Systems Fail?, 2025. URL https://arxiv.org/abs/2503.13657.

Kang, Z., Zhao, X., and Song, D. Scalable best-of-n selection for large language models via self-certainty. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=29FRqmVQK8.

Kazeev, N. and Ustyuzhanin, A. E. Uncertainty quantification for LLM agents via semantic abstraction trajectories. In *ICML 2026 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2026. URL https://openreview.net/forum?id=ZsThsDtfaA.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. In Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., and Sun, Y. (eds.), *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Liu, A., Mei, A., Lin, B., Xue, B., Wang, B., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., et al. DeepSeek-V3.2:

Pushing the Frontier of Open Large Language Models. *arXiv preprint arXiv:2512.02556*, 2025.

Liu, X., Yang, X., Li, Z., Li, P., and He, R. AgentHallu: Benchmarking Automated Hallucination Attribution of LLM-based Agents, 2026. URL https://arxiv.org/abs/2601.06818.

Luce, R. D. et al. *Individual choice behavior*, volume 4. Wiley New York, 1959.

Marks, S. and Tegmark, M. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Plackett, R. L. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24 (2):193–202, 1975.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., and Barzilay, R. Conformal Language Modeling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pzUhfQ74c5.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Sarfati, R., Liu, T. J. B., Boullé, N., and Earls, C. J. Lines of thought in large language models, 2025. URL https://arxiv.org/abs/2410.01545.

Stechly, K., Valmeekam, K., and Kambhampati, S. Chain of Thoughtlessness? An Analysis of CoT in Planning. *Advances in Neural Information Processing Systems*, 37: 29106–29141, 2024.

Wang, M., Lin, R., Hu, K., Jiao, J., Chowdhury, N., Chang, E., and Patwardhan, T. FrontierScience: Evaluating AI's Ability to Perform Expert-Level Scientific Tasks. *arXiv preprint arXiv:2601.21165*, 2026.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Weng, Y., Zhu, M., Xia, F., Li, B., He, S., Liu, S., Sun, B., Liu, K., and Zhao, J. Large Language Models are Better Reasoners with Self-Verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. URL https://aclanthology.org/2023.findings-emnlp.167/.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Zhang, J., Xiong, C., and Wu, C.-S. Agentic confidence calibration, 2026. URL https://arxiv.org/abs/2601.15778.

Zhao, Q., Li, D., Liu, Y., Cheng, W., Sun, Y., Oishi, M., Osaki, T., Matsuda, K., Yao, H., Zhao, C., Chen, H., and Zhao, X. Uncertainty propagation on LLM agent. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6064–6073, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.302. URL https://aclanthology.org/2025.acl-long.302/.

Zhao, Z., Koishekenov, Y., Yang, X., Murray, N., and Cancedda, N. Verifying Chain-of-Thought Reasoning via Its Computational Graph, 2026. URL https://arxiv.org/abs/2510.09312.

Zhu, W., Zhang, J., Yu, L., Yue, K., and Tang, Z. Dissecting Failure Dynamics in Large Language Model Reasoning, 2026. URL https://arxiv.org/abs/2604.14528.