Spectral Coherence for Intent Commitment in Tool-Using Language Agents

Anonymous ACL submission

Abstract

Tool-using language agents often fail before execution by committing to an underspecified or invalid intent. We propose *spectral coherence*, a hidden-state diagnostic for intent-hypothesis verification. The premise is that a supported hypothesis organizes token representations around shared goal directions, producing covariance outliers beyond an unstructured random-matrix bulk. From a single hypothesis-conditioned forward pass, spectral coherence measures this concentration relative to a Marchenko–Pastur reference, providing a signal distinct from log-probability or verbalized confidence. On τ-bench retail/airline, τ^2-bench telecom, and ToolBench, using Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3, spectral coherence improves correct-hypothesis AUROC over length-normalized log-probability by 6.5–9.8 points. Null controls indicate that the signal tracks user-hypothesis alignment rather than length, specificity, JSON format, or generic policy structure. In a minimal commit-or-clarify wrapper with feasibility checks, it improves task reward by 5.8–8.2 points and reduces policy violations by 27–34.7% relative to generation.

1 Introduction

Tool-using language agents often fail before tool execution. A user request to change an order, update a flight, or reconfigure a service account can support multiple intent hypotheses, some incomplete, ambiguous, or policy-invalid. Before selecting a tool, the agent must decide whether one hypothesis is sufficiently supported by the user message, or whether it should ask for clarification.

Standard confidence signals are indirect for this decision. Token probabilities, entropy, self-consistency, and verbalized confidence measure output behavior, but do not test whether model has internally organized the context around a coherent intent hypothesis (Qin et al., 2024a; Hu et al., 2026; Liu et al., 2024). A high-likelihood continuation can still commit to the wrong intent, while a clarification-appropriate hypothesis may be disfavored by next-token likelihood.

We ask whether intent support leaves a *measurable geometric trace* in hidden states. Our hypothesis is that a supported intent aligns token representations around shared goal directions induced by the user, dialogue, policy, and candidate hypothesis. By contrast, an irrelevant, underspecified, or inconsistent hypothesis should induce weaker shared structure. We measure this effect with *spectral coherence*: the token-token covariance mass that lies outside the Marchenko–Pastur bulk of an unstructured random-matrix reference (Yaskov, 2023).

Spectral coherence is a comparative verification signal that measures whether a candidate hypothesis induces coherent goal structure, but not whether the associated action is complete, allowed, or semantically valid. We therefore use it to rank competing hypotheses for the same user context and combine it with feasibility and policy checks before action commitment. We evaluate it as an intent-hypothesis ranker on τ-bench retail/airline (Yao et al., 2024), τ^2-bench telecom (Barres et al., 2025), and ToolBench (Qin et al., 2024b; Guo et al., 2024); use null controls for length, specificity, JSON format, policy-only structure, and paraphrase; and test downstream utility in a lightweight commit-or-clarify wrapper. Across three tool-use benchmarks and three open-weight model families, spectral coherence improves hypothesis ranking over length-normalized log-probability, tracks user-hypothesis alignment rather than surface artifacts, and improves a lightweight commit-or-clarify wrapper in both reward and policy-violation rate.

2 Related Work

Tool-using agents and benchmarks. ReAct interleaves reasoning and acting for language

agents (Yao et al., 2022). ToolBench covers API selection over 16,464 real-world APIs (Qin et al., 2024b). τ-bench evaluates agent–user conversations in retail and airline domains (Yao et al., 2024), and τ^2-bench extends this to dual-control telecom tasks (Barres et al., 2025). Our work targets the internal verification before tool commitment.

Uncertainty and verification for LLMs. Self-consistency (Wang et al., 2022), token entropy, and verbalized confidence are useful but remain output-space signals (Shorinwa et al., 2025; Liu et al., 2024). Our method measures hidden-state geometry induced by candidate hypotheses, complementing feasibility checks and conformal commit/clarify decisions.

Random matrix theory and neural representations. RMT has been applied to neural weight spectra, generalization, heavy-tailed self-regularization, and out-of-distribution detection. We use the Marchenko–Pastur law as a diagnostic reference for per-forward-pass hidden-state covariances conditioned on a specific user context and hypothesis (Tao, 2012; Staats et al., 2026; Darabi et al., 2026; Dantas et al., 2025; Ettori et al., 2026).

3 Problem Setup

At dialogue turn t, a tool-using agent observes a system policy π, dialogue history h_t, user message u_t, and tool schemas $\mathcal{A}$. Before acting, the agent must decide whether the context supports a concrete intent or whether the user should be asked for clarification. We formulate this as hypothesis verification. A hypothesis generator proposes K candidate intent hypotheses, $\mathcal{Z}_t = \{z_{t,1}, \ldots, z_{t,K}\}$, where each $z_{t,k}$ specifies an inferred goal, missing information, relevant policy constraints, and an action type such as `tool_call`, `ask_clarification`, `final_response`, or `refuse_or_redirect`. The verifier must rank these hypotheses and select either a commit hypothesis, which permits tool execution, or a clarification hypothesis, which defers action until required information is available.

For ranking experiments, one hypothesis is labeled correct or clarification-appropriate according to the ground-truth task state. For agent experiments, the selected hypothesis determines the next environment action. We evaluate hypothesis-ranking AUROC and downstream task reward, pass^k, clarification F1, and policy violation rate.

4 Spectral Coherence

For a layer ℓ, let $\mathbf{H}^{(\ell)} \in \mathbb{R}^{n\times d}$ denote the hidden states of the non-padding tokens in a hypothesis-augmented prompt. We center over tokens,

$$\tilde{\mathbf{H}}^{(\ell)} = \mathbf{H}^{(\ell)} - \mathbf{1}\bar{\mathbf{h}}^{(\ell)\top}, \qquad \bar{\mathbf{h}}^{(\ell)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{H}^{(\ell)}_{i,:}, \tag{1}$$

and form the token-token covariance

$$\mathbf{C}^{(\ell)} = \frac{1}{d}\tilde{\mathbf{H}}^{(\ell)}\tilde{\mathbf{H}}^{(\ell)\top} \in \mathbb{R}^{n\times n}. \tag{2}$$

Let $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$ be its eigenvalues.

Spectral coherence measures covariance mass outside an unstructured reference bulk. If the centered token representations were independent isotropic Gaussians with variance σ^2, the eigenvalue distribution of $\mathbf{C}^{(\ell)}$ would follow the Marchenko–Pastur (MP) law with aspect ratio $q = n/d$ and upper edge $\lambda^+ = \sigma^2(1+\sqrt{q})^2$.

Transformer hidden states do not satisfy this null model, so we use the MP edge only as a diagnostic reference for paired comparisons among hypotheses with the same context, prompt template, and model. We estimate the bulk scale by median matching, $\hat{\sigma}^2 = \mathrm{median}(\lambda_i)/m_q$, where m_q is the median of the unit-variance MP distribution for aspect ratio q. For each layer, we compute

$$\textsc{MPDev}^{(\ell)} = \sum_i \max(0, \lambda_i - \lambda^+), \tag{3}$$

$$\textsc{OutMass}^{(\ell)} = \frac{\sum_{i:\lambda_i>\lambda^+} \lambda_i}{\sum_i \lambda_i}, \tag{4}$$

$$\textsc{SpEntropy}^{(\ell)} = -\sum_i p_i \log p_i, p_i = \lambda_i / \sum_j \lambda_j. \tag{5}$$

Features are min-max normalized across the K hypotheses for each turn. The layer score is

$$s_\ell(z_k) = w_1 \widehat{\textsc{MPDev}}^{(\ell)}(z_k) + w_2 \widehat{\textsc{OutMass}}^{(\ell)}(z_k) - w_3 \widehat{\textsc{SpEntropy}}^{(\ell)}(z_k), \tag{6}$$

with $(w_1, w_2, w_3) = (0.45, 0.35, 0.20)$ selected on calibration tasks. The final coherence score averages over selected layers:

$$s_{\mathrm{coh}}(z_k) = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}} s_\ell(z_k). \tag{7}$$

The intended signal is goal-aligned concentration. A supported intent hypothesis should align

Table 1: Models and hidden layers used for spectral extraction. Layer indices are zero-based block indices.

Model	Blocks	$\mathcal{L}$
Qwen2.5-7B-Instruct	28	{7, 14, 21, 27}
Llama-3.1-8B-Instruct	32	{8, 16, 24, 31}
Mistral-7B-Instruct-v0.3	32	{8, 16, 24, 31}

Table 2: Spectral feature gaps, correct minus alternative, on 500 τ-bench probing tasks. Standard errors are in parentheses. *: paired $p < 0.01$ after Holm correction.

Condition	ΔMPD	ΔOM	ΔSE
Plaus.-wrong*	.113 (.018)	.047 (.007)	-.061 (.010)
Matched-spec.*	.096 (.016)	.041 (.006)	-.052 (.009)
Shuffled-user*	.151 (.020)	.063 (.008)	-.078 (.011)
Struct.-irr.*	.128 (.019)	.052 (.007)	-.069 (.010)
Policy-only*	.082 (.015)	.034 (.006)	-.044 (.009)
Paraphrase	.006 (.007)	.003 (.003)	-.004 (.004)

token representations around shared directions induced by the user, dialogue, policy, and candidate hypothesis, increasing out-of-bulk mass and reducing spectral entropy. Unsupported or inconsistent hypotheses should induce weaker shared structure.

5 Controls and Hypothesis Ranking

We first test whether spectral coherence reflects user-hypothesis alignment rather than surface artifacts. For each probing task, we construct one correct or clarification-appropriate hypothesis and five alternatives: plausible-wrong, matched-specificity wrong, shuffled-user, structured-irrelevant, policy-only, and paraphrase. The alternatives preserve different surface properties while breaking intent alignment: plausible-wrong changes a key argument; matched-specificity matches token count, entity count, and policy-reference count; shuffled-user preserves user tokens but destroys order; structured-irrelevant uses a valid goal from another task; policy-only preserves policy structure without user intent; and paraphrase preserves the correct semantics while changing surface form. Incorrect hypotheses are produced externally rather than by the evaluated model to reduce contamination.

Table 2 shows that the signal survives matched-specificity and policy-only controls, is largest when user-context alignment is disrupted, and is near zero under paraphrase. This supports the interpretation that spectral coherence tracks semantic alignment between the user context and candidate hypothesis rather than verbosity, formatting, or generic policy structure.

Table 3: Standalone hypothesis-ranking AUROC. Full results with standard errors are in Appendix A.

Model	Ranker	τ-R	τ-A	τ^2-T	TB
Qwen	Uniform	.500	.500	.500	.500
	Logprob	.714	.692	.668	.701
	Spectral	**.812**	**.781**	**.754**	**.793**
Llama	Uniform	.500	.500	.500	.500
	Logprob	.697	.674	.653	.682
	Spectral	**.774**	**.752**	**.725**	**.759**
Mistral	Uniform	.500	.500	.500	.500
	Logprob	.681	.661	.641	.669
	Spectral	**.751**	**.733**	**.706**	**.741**

We then evaluate spectral coherence as a standalone ranker on τ-bench retail/airline, τ^2-bench telecom, and ToolBench solution-path API selection. For each task, the ranker receives $K = 5$ hypotheses with one labeled correct or clarification-appropriate. We report AUROC for assigning the labeled hypothesis a higher score than alternatives. The spectral ranker uses s_{coh}; the logprob uses length-normalized hypothesis log-probability.

Spectral coherence improves over length-normalized log-probability for every model and benchmark in Table 3. Gains are largest on τ-bench retail, where policy and entity structure provide strong alignment cues, and smallest on τ^2 telecom, where multiple hypotheses can be internally coherent while only one is operationally feasible. This pattern is consistent with the intended role of spectral coherence as a verification signal rather than a complete policy-validity test.

6 Commit-or-Clarify Wrapper

To test whether spectral coherence is useful beyond standalone ranking, we insert it into a lightweight commit-or-clarify wrapper. At each turn, the wrapper generates $K = 5$ candidate hypotheses, including at least one clarification hypothesis. Each hypothesis is checked for tool availability, argument validity, missing required slots, policy compliance, authentication requirements, and sensitive-action confirmation. The hypotheses are then scored with a calibrated combination of length-normalized log-probability, spectral coherence, feasibility, attention alignment to user/policy/tool spans, missing-slot density, and contradiction count.

The wrapper commits to a tool action only when the top hypothesis is feasible and the calibrated prediction set is singleton; otherwise, it asks the highest-ranked clarification question. The calibra-

Table 4: Commit-or-clarify. Reward/pass^1/violation are averaged over five trials.

Model	Domain/System	Reward	pass^1	Viol.
Qwen	Retail Direct	.612	.548	.118
	Retail Hyp-only	.659	.598	.104
	Retail SpecHyp	**.694**	**.641**	**.077**
	Airline Direct	.571	.504	.134
	Airline Hyp-only	.615	.552	.121
	Airline SpecHyp	**.652**	**.603**	**.089**
	Telecom Direct	.438	.392	.161
	Telecom Hyp-only	.471	.425	.149
	Telecom SpecHyp	**.518**	**.472**	**.108**
Llama	Retail Direct	.589	.527	.126
	Retail Hyp-only	.632	.572	.112
	Retail SpecHyp	**.663**	**.613**	**.086**
	Airline Direct	.552	.487	.142
	Airline Hyp-only	.592	.529	.128
	Airline SpecHyp	**.626**	**.578**	**.098**
	Telecom Direct	.419	.371	.169
	Telecom Hyp-only	.451	.407	.157
	Telecom SpecHyp	**.489**	**.444**	**.119**
Mistral	Retail Direct	.566	.501	.137
	Retail Hyp-only	.604	.544	.121
	Retail SpecHyp	**.631**	**.585**	**.095**
	Airline Direct	.529	.463	.151
	Airline Hyp-only	.567	.508	.136
	Airline SpecHyp	**.599**	**.553**	**.107**
	Telecom Direct	.401	.354	.181
	Telecom Hyp-only	.429	.386	.168
	Telecom SpecHyp	**.459**	**.419**	**.132**

tion rule is turn-level and marginal, not a trajectory-level guarantee. Its role is to make the downstream experiment conservative: spectral coherence can support commitment, but feasibility and calibration decide whether commitment is allowed.

7 Evaluation

Setup. We evaluate the wrapper on τ-bench retail/airline and τ^2-bench telecom. ToolBench is used only for standalone ranking because it lacks the same interactive commit-or-clarify environment. We compare direct generation, hypothesis-only log-probability ranking, and the full spectral wrapper. Each task is evaluated with five trials. We report mean reward, pass^1, and policy violation rate; standard errors, additional baselines, clarification metrics, and implementation details are in the appendix.

Table 4 shows consistent gains across model families and domains. For Qwen2.5-7B, SpecHyp improves direct generation by 8.2 reward points on retail, 8.1 on airline, and 8.0 on telecom, while reducing violations by 34.7%, 33.6%, and 32.9%. Gains are smaller but consistent for Llama-3.1-8B and Mistral-7B. Improvements over hypothesis-only ranking show that spectral coherence adds verification signal beyond explicit hypothesis generation.

Table 5: Ablation on Qwen2.5-7B, τ-bench retail.

Variant	Reward	pass^1	Viol.
Direct	.612	.548	.118
Hyp-only	.659	.598	.104
No spectral	.675	.622	.081
Random spectral	.669	.615	.085
No feasibility	.676	.623	.132
Full SpecHyp	**.694**	**.641**	**.077**

Clarification, ablation, and confounds. The wrapper asks slightly more questions, increasing average turn count by 0.7–0.8 turns on Qwen2.5-7B, but they are more targeted: clarification F1 improves from .392 to .611 on retail, .368 to .584 on airline, and .331 to .557 on telecom. Table 5 isolates the spectral contribution. Removing spectral coherence lowers reward from .694 to .675, and random spectral scores give .669. Removing feasibility increases violations from .077 to .132, showing that spectral coherence helps rank commitment candidates but cannot replace policy and argument checks. Confound checks support the same interpretation: length-normalized MP deviation, equal-length padding, and JSON field-order shuffling change AUROC by at most .006, while removing user-history tokens drops AUROC from .812 to .704.

Mechanistic check. Steering and hidden-state swap analyses suggest that the measured spectral subspaces are decision-relevant. Moderate contrastive steering raises Qwen2.5-7B retail reward from .694 to .703, random projectors do not improve reward, and counterfactual swaps change the commit/clarify decision in 31.8% of failed trajectories. These results are suggestive, not a full mediation claim.

8 Conclusion

Spectral coherence is a hidden-state diagnostic for intent-hypothesis verification in tool-using agents. Across τ-bench, τ^2-bench, and ToolBench, it improves hypothesis ranking over length-normalized log-probability, survives surface-form controls, and improves a lightweight commit-or-clarify wrapper when paired with feasibility checks. The conclusion is conservative: hidden-state spectral geometry provides useful evidence that an agent has formed a coherent goal representation, but correctness still requires policy and environment-level verification.

Limitations

Spectral coherence requires access to hidden states and is therefore not directly applicable to closed-source APIs without model-side instrumentation. The Marchenko–Pastur reference is a diagnostic abstraction, not a literal model of transformer hidden states; the empirical claims rely on paired comparisons among hypotheses sharing the same context, prompt template, and model. The method is also not a correctness guarantee. A model can form a coherent representation of an invalid or unsafe goal, so feasibility, policy, and environment-state checks remain necessary. Our evaluation is limited to structured tool-use benchmarks with explicit policies, schemas, and candidate hypotheses. Performance may differ in open-ended dialogue, long-horizon multi-agent settings, or domains where hypothesis construction is poor. The conformal commit-or-clarify rule is turn-level and marginal; it does not provide a trajectory-level safety guarantee and may require recalibration under domain shift. Finally, spectral extraction adds computation because multiple hypotheses must be evaluated and eigenspectra computed, although batching and selected-layer extraction reduce this overhead.

Ethics Statement

This work studies reliability diagnostics for tool-using language agents. The method may reduce unsafe or premature tool calls by encouraging clarification under uncertainty, but it can also create false confidence if spectral coherence is misinterpreted as correctness. We therefore pair spectral features with feasibility checks and emphasize the diagnostic nature of the signal.

References

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. 2025. τ^2-bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*.

Pierre Dantas, Waldir Junior, Lucas Cordeiro, and Eulanda Santos. 2025. Decoding transformers spectra: A random matrix theory framework beyond the marchenko–pastur law. *ResearchSquare preprint*.

Nastaran Darabi, Devashri Naik, Sina Tayebati, Dinithi Jayasuriya, Ranganath Krishnan, and Amit Ranjan Trivedi. 2026. Eigenshield: Inference-time, model-agnostic jailbreaking defense via causal subspace filtering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3524–3532.

Davide Ettori, Nastaran Darabi, Sureshkumar Senthilkumar, and Amit Ranjan Trivedi. 2026. Rmt-kd: Random matrix theoretic causal knowledge distillation. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6181–6185. IEEE.

Zhicheng Guo, Sijie Cheng, Hao Wang, Shihao Liang, Yujia Qin, Peng Li, Zhiyuan Liu, Maosong Sun, and Yang Liu. 2024. Stabletoolbench: Towards stable large-scale benchmarking on tool learning of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11143–11156.

Jinchao Hu, Meizhi Zhong, Kehai Chen, Xuefeng Bai, and Min Zhang. 2026. Agentic tool use in large language models. *arXiv preprint arXiv:2604.00835*.

Hao Liu, Zi-Yi Dou, Yixin Wang, Nanyun Peng, and Yisong Yue. 2024. Uncertainty calibration for tool-using language agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16781–16805.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, and 1 others. 2024a. Tool learning with foundation models. *ACM Computing Surveys*, 57(4):1–40.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, and 1 others. 2024b. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pages 9695–9717.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 58(3):1–38.

Max Staats, Matthias Thamm, and Bernd Rosenow. 2026. Small singular values matter: A random matrix analysis of transformer models. *Advances in Neural Information Processing Systems*, 38:153545–153573.

Terence Tao. 2012. *Topics in random matrix theory*, volume 132. American Mathematical Society.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Pavel Yaskov. 2023. Marchenko-pastur law for a random tensor model. *Electronic Communications in Probability*, 28:1–17.

A Full Standalone Ranking Results

Table 6 reports the full standalone hypothesis-ranking AUROC results with standard errors. These are the expanded values corresponding to Table 3 in the main text.

B Wrapper Scoring and Calibration Details

At each turn, the wrapper scores candidate hypotheses using spectral coherence together with features that capture complementary failure modes. For hypothesis z_k, the score is

$$\begin{aligned} S(z_k) = \alpha_0 \hat{p}_k + \beta_0 \hat{s}^{(k)}_{\text{coh}} + \gamma_0 \hat{s}^{(k)}_{\text{user}} \\ + \rho_0 \hat{s}^{(k)}_{\text{policy}} + \eta_0 \hat{s}^{(k)}_{\text{tool}} + \omega_0 \hat{f}^{(k)} \\ + \nu_0 \hat{\nu}^{(k)} - \delta_0 \hat{d}^{(k)} - \kappa_0 \hat{c}^{(k)}. \end{aligned} \tag{8}$$

Here $\hat{p}_k$ is length-normalized hypothesis log-probability, $\hat{s}^{(k)}_{\text{coh}}$ is spectral coherence, $\hat{s}^{(k)}_{\text{user}}$, $\hat{s}^{(k)}_{\text{policy}}$, and $\hat{s}^{(k)}_{\text{tool}}$ are attention-alignment scores to user, policy, and tool-schema spans, $\hat{f}^{(k)}$ is feasibility, $\hat{\nu}^{(k)}$ is eigenspace stability, $\hat{d}^{(k)}$ is missing-slot density, and $\hat{c}^{(k)}$ is contradiction count against the dialogue state. We use

$$\begin{aligned} (\alpha_0, \beta_0, \gamma_0, \rho_0, \eta_0, \omega_0, \nu_0, \delta_0, \kappa_0) \\ = (0.18,\ 0.24,\ 0.13,\ 0.12,\ 0.08, \\ 0.20,\ 0.07,\ 0.10,\ 0.18), \end{aligned} \tag{9}$$

selected on a disjoint calibration split.

Scores are softmax-normalized into $P(z_k)$. On calibration turns, the nonconformity score is

$$R_i = 1 - P(z_i^*), \tag{10}$$

where z_i^* is the labeled correct or clarification-appropriate hypothesis. Let $\hat{q}_{0.90}$ be the finite-sample split-conformal quantile. The prediction set is

$$\mathcal{C}(x) = \{z_k : 1 - P(z_k) \leq \hat{q}_{0.90}\}. \tag{11}$$

The wrapper commits only when $|\mathcal{C}(x)| = 1$ and $f(z_{k^*}) > 0.70$; otherwise it asks the highest-ranked clarification question. The guarantee is turn-level and marginal under exchangeability. It is not a trajectory-level safety guarantee and may require recalibration under domain shift.

C Full Agent Results

Table 7 expands the compact commit-or-clarify results in the main text with additional baselines and the spectral-steering variant.

D Clarification and Reliability

Table 8 reports pass^2, pass^4, clarification F1, and average turns for Qwen2.5-7B-Instruct. SpecHyp-Agent asks slightly more questions, but the additional turns are targeted: clarification F1 improves substantially across all domains.

E Ablation Study

Table 9 gives the full ablation on Qwen2.5-7B-Instruct, τ-bench retail. Hypothesis generation alone improves over direct generation, but adding spectral coherence improves reward further. Random spectral scores do not recover the full gain. Removing feasibility substantially increases policy violations, confirming that spectral coherence must be paired with policy and argument checks.

F Confound Analysis

Table 10 tests whether the spectral signal is explained by simple artifacts. Length normalization, equal-length padding, and JSON field-order shuffling have small effects. Removing user-history tokens causes the largest degradation, indicating that spectral coherence is not merely a property of the hypothesis text. Final-layer-only extraction also degrades performance, consistent with the layer-localization results.

G Spectral Steering and Layer Analysis

Spectral Steering. We compute success and failure covariance averages on calibration trajectories, extract top-r eigenspaces, and apply the forward-hook intervention

$$\mathbf{H}^{(\ell)} \leftarrow \mathbf{H}^{(\ell)} - \alpha \mathbf{P}_- \mathbf{H}^{(\ell)} + \alpha \mathbf{P}_+ \mathbf{H}^{(\ell)}. \tag{12}$$

Here $\mathbf{P}_+$ and $\mathbf{P}_-$ are projectors onto success- and failure-associated subspaces. We compare against random orthogonal projectors of the same rank.

Moderate contrastive steering improves reward, while large steering degrades performance. Random projectors do not improve reward and become harmful at larger strengths.

Layer Localization. Mid-layer interventions produce larger reward changes than early or final-layer interventions, suggesting that the relevant

Table 6: Standalone hypothesis-ranking AUROC. Values are means over three seeds; standard errors are in parentheses.

Model	Ranker	τ-bench Retail	τ-bench Airline	τ^2 Telecom	ToolBench
Qwen2.5-7B-Instruct	Uniform	0.500	0.500	0.500	0.500
	Logprob	0.714 (0.008)	0.692 (0.010)	0.668 (0.011)	0.701 (0.009)
	Spectral	**0.812 (0.007)**	**0.781 (0.009)**	**0.754 (0.010)**	**0.793 (0.008)**
Llama-3.1-8B-Instruct	Uniform	0.500	0.500	0.500	0.500
	Logprob	0.697 (0.010)	0.674 (0.011)	0.653 (0.012)	0.682 (0.010)
	Spectral	**0.774 (0.009)**	**0.752 (0.010)**	**0.725 (0.011)**	**0.759 (0.009)**
Mistral-7B-Instruct-v0.3	Uniform	0.500	0.500	0.500	0.500
	Logprob	0.681 (0.011)	0.661 (0.012)	0.641 (0.012)	0.669 (0.011)
	Spectral	**0.751 (0.010)**	**0.733 (0.011)**	**0.706 (0.012)**	**0.741 (0.010)**

Table 7: Agent results on τ-bench retail/airline and τ^2-bench telecom. Rewards are means over five trials. Best non-steering result per model/domain is bolded.

		τ-bench Retail			τ-bench Airline			τ^2 Telecom		
Model	System	Reward	pass^1	Viol.	Reward	pass^1	Viol.	Reward	pass^1	Viol.
Qwen2.5-7B	Direct	0.612	0.548	0.118	0.571	0.504	0.134	0.438	0.392	0.161
	Self-consistency	0.648	0.586	0.109	0.603	0.541	0.126	0.462	0.417	0.155
	Token entropy	0.627	0.563	0.115	0.586	0.522	0.131	0.447	0.401	0.159
	Hyp-only	0.659	0.598	0.104	0.615	0.552	0.121	0.471	0.425	0.149
	SpecHyp-Agent	**0.694**	**0.641**	**0.077**	**0.652**	**0.603**	**0.089**	**0.518**	**0.472**	**0.108**
	+Spectral steering	0.703	0.649	0.076	0.660	0.611	0.088	0.529	0.483	0.105
Llama-3.1-8B	Direct	0.589	0.527	0.126	0.552	0.487	0.142	0.419	0.371	0.169
	Hyp-only	0.632	0.572	0.112	0.592	0.529	0.128	0.451	0.407	0.157
	SpecHyp-Agent	**0.663**	**0.613**	**0.086**	**0.626**	**0.578**	**0.098**	**0.489**	**0.444**	**0.119**
	+Spectral steering	0.670	0.620	0.085	0.634	0.585	0.096	0.497	0.453	0.116
Mistral-7B	Direct	0.566	0.501	0.137	0.529	0.463	0.151	0.401	0.354	0.181
	Hyp-only	0.604	0.544	0.121	0.567	0.508	0.136	0.429	0.386	0.168
	SpecHyp-Agent	**0.631**	**0.585**	**0.095**	**0.599**	**0.553**	**0.107**	**0.459**	**0.419**	**0.132**
	+Spectral steering	0.638	0.591	0.094	0.606	0.559	0.106	0.466	0.426	0.130

Table 8: Reliability and clarification metrics on Qwen2.5-7B-Instruct.

Domain/System	pass^2	pass^4	Clar. F1	Avg. turns
Retail Direct	0.684	0.775	0.392	7.1
Retail SpecHyp	**0.759**	**0.836**	**0.611**	7.8
Airline Direct	0.641	0.742	0.368	7.5
Airline SpecHyp	**0.721**	**0.811**	**0.584**	8.2
Telecom Direct	0.507	0.612	0.331	9.3
Telecom SpecHyp	**0.598**	**0.697**	**0.557**	10.1

Table 9: Ablation on Qwen2.5-7B-Instruct, τ-bench retail.

Variant	Reward	pass^1	Viol.
Direct generation	0.612	0.548	0.118
Hyp-only logprob	0.659	0.598	0.104
Hyp + spectral only	0.672	0.616	0.101
Hyp + feasibility	0.671	0.617	0.083
Logprob + spectral	0.681	0.628	0.097
No-Spectral	0.675	0.622	0.081
No-Feasibility	0.676	0.623	0.132
No-Clarify	0.661	0.609	0.119
No-Stability	0.686	0.634	0.081
Random spectral score	0.669	0.615	0.085
Uniform weights	0.681	0.629	0.088
Full SpecHyp-Agent	**0.694**	**0.641**	**0.077**

goal structure is most accessible in intermediate representations.

Counterfactual Hidden-State Swaps. On failed Qwen2.5 retail trajectories where the selected hypothesis was not the labeled correct hypothesis, replacing selected-hypothesis hidden states with correct-hypothesis hidden states changed the commit/clarify decision in 31.8% of cases and changed it to the labeled correct decision in 22.6% of cases. The reverse swap, replacing correct-hypothesis states with plausible-wrong states, reduced correct selection by 18.9%. These swaps do not prove full mediation, but they show that spectral differences are often large enough to affect the decision boundary.

Table 10: Confound checks on Qwen2.5-7B-Instruct, τ-bench retail.

Check	AUROC	Change
Main spectral score	0.812	–
Length-normalized MPDev	0.806	-0.006
Equal-length padding	0.808	-0.004
JSON field order shuffled	0.809	-0.003
Final layer only	0.753	-0.059
No policy tokens	0.781	-0.031
No user-history tokens	0.704	-0.108

Table 11: Spectral steering sweep on Qwen2.5-7B-Instruct, τ-bench retail.

α	Contrastive steering reward	Random steering reward
0.00	0.694	0.694
0.05	0.700	0.693
0.10	**0.703**	0.692
0.15	0.699	0.690
0.20	0.689	0.684
0.30	0.662	0.671

H Effective Rank Ablation

Effective rank gives little additional signal because it is closely related to spectral entropy. Table 13 shows that adding effective rank produces only minor AUROC changes beyond MPDev, OutMass, and SpEntropy.

I Spiked-Covariance Motivation

The MP reference in the main text is used only as a diagnostic null. A simple spiked-covariance model motivates why goal-aligned hidden-state structure may produce out-of-bulk eigenvalues.

Proposition 1 (Idealized goal-spike separation)**.** *Consider an idealized hidden-state model*

$$\tilde{\mathbf{H}} = \mathbf{Z} + \sqrt{d\rho}\,\mathbf{a}\mathbf{b}^\top, \tag{13}$$

where $\mathbf{Z}_{ij} \sim \mathcal{N}(0, \sigma^2)$*,* $\|\mathbf{a}\|_2 = \|\mathbf{b}\|_2 = 1$*, and* $q = n/d$ *has a non-zero limit. In the associated spiked covariance model, a sufficiently strong rank-one goal component produces a sample eigenvalue outside the MP bulk, while a subcritical component remains absorbed into the bulk. The critical threshold is governed by the standard BBP phase transition for spiked covariance matrices.*

This proposition is only an intuition for the diagnostic. Transformer hidden states do not satisfy the independence and isotropy assumptions of the model. In our setting, $\mathbf{a}\mathbf{b}^\top$ abstracts a shared goal direction induced by the user, dialogue, policy, and candidate hypothesis. The empirical tests determine if abstraction is useful for verification.

Table 12: Layer-specific steering on Qwen2.5-7B-Instruct, τ-bench retail, $\alpha = 0.10$.

Steered layer(s)	ΔReward	Δs_{coh}
Layer 7 only	+0.004	+0.013
Layer 14 only	+0.011	+0.031
Layer 21 only	+0.014	+0.036
Layer 27 only	+0.006	+0.018
All selected layers	+0.009	+0.044

Table 13: Effective-rank ablation on Qwen2.5-7B-Instruct.

Features	Retail	Airline	Telecom	ToolBench
MPDev + OutMass	0.803	0.772	0.746	0.784
+ SpEntropy	0.812	0.781	0.754	0.793
+ EffRank	0.814	0.782	0.756	0.795

J Computational Cost

SpecHyp-Agent adds overhead because it processes multiple hypotheses and computes eigenspectra. In our implementation, the median per-turn latency multiplier relative to direct generation is 2.7× for Qwen2.5-7B, 2.9× for Llama-3.1-8B, and 2.6× for Mistral-7B. Batching hypothesis prompts and extracting only selected layers keeps the cost below a naive K-forward-pass implementation. The cost can be reduced further by computing spectra only for a subset of layers or by using approximate eigensolvers for long contexts.

K Implementation Details

Hidden states are captured with forward hooks and removed immediately after generation. Hypothesis prompts are batched with left padding, and covariance is computed only over non-padding tokens. Eigendecomposition uses double precision. For numerical stability, spectra with total mass below 10^{-8} are skipped and the turn is assigned the median calibration score. Random seeds are fixed for Python, NumPy, and PyTorch at trial start.

All features used in the wrapper are computed per turn and normalized across the $K = 5$ candidate hypotheses for the same user context. This paired normalization ensures that scores compare hypotheses within a shared prompt and model context rather than across unrelated turns. Spectral variants use eager attention to expose attention weights when alignment features are required; direct generation uses optimized scaled-dot-product attention when attention weights are not needed.