Knowing When to Ask: Self-Gated Clarification for Hierarchical Language Agents

Aijing Gao, Yiming Kang, Mengdie Flora Wang, Jae Oh Woo

Amazon Web Services
{gaijing, ymkang, florawan, jaeohwoo}@amazon.com

Abstract

In hierarchical reasoning, failures often originate at intermediate decision points where the agent commits to a wrong branch without recognizing that it lacks critical information. We propose ACTIONRATING, a formulation that places clarification inside the agent's action space on a shared ordinal scale with navigation, so that asking competes directly with acting at every decision point. Two structurally distinct information-seeking modes emerge from the agent's own ratings: *mandatory* (no viable branch) and *opportunistic* (residual uncertainty despite a leading candidate). On Harmonized Tariff Schedule classification (30,000-node taxonomy, three benchmarks, 9 LLMs across 4 families), we observe a regime shift from mandatory to opportunistic clarification, with Information-Seeking Effectiveness (ISE), the fraction of help interactions followed by a correct next step, rising from 50% to 74%. Three diagnostic contrasts fail to reproduce this structure. A separability test confirms that the information-seeking pattern persists even when answer quality is degraded (-18.8% accuracy), demonstrating that *where* an agent seeks help can be analyzed independently of *what* help it receives. Under controlled answers, accuracy gains reach $+16.2\%$ at 10-digit, bounding what better localization could unlock.

1 Introduction

Language agents that reason over hierarchical structures, medical codes, legal statutes, product taxonomies, face a recurring failure mode: once the agent commits to a wrong intermediate branch, every subsequent step merely elaborates an error that should have been caught earlier [Yao et al., 2023b,a, Shinn et al., 2023, Press et al., 2023, Dziri et al., 2024]. Final-answer accuracy tells us *that* the system failed, but not *where*, at which decision point the agent lacked the information to proceed safely. The core question is deceptively simple: *when should the agent ask for help instead of committing?*

Why current approaches fall short. Existing designs treat clarification as external to the reasoning trajectory: a confidence threshold [Kadavath et al., 2022], a prompt instruction ("ask if unsure"), or sampling-based disagreement [Kuhn et al., 2023]. These mechanisms decouple the decision to ask from the decision to act, leaving two problems unsolved. First, they do not make information-seeking behavior *structurally observable*, we cannot distinguish an agent that asks because no viable branch exists from one that asks to reduce residual uncertainty. Second, they confound *where* help was sought with *what* help was received: an agent that asks more may perform better simply because it gets better information.

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: AI Agents for Discovery in the Wild.

Clarification as action. We propose ACTIONRATING, a formulation that addresses both problems by placing clarification inside the agent's own action space (Figure 1). The agent scores candidate next actions, including a dedicated clarification action, on a shared $[0, 100]$ ordinal scale, so that asking competes directly with acting at each decision point. This shared-scale competition makes the local need for help observable without any external uncertainty estimator. Two structurally distinct modes emerge from the agent's own ratings: *mandatory* help, where clarification is top-ranked and no navigation branch is viable, and *opportunistic* help, where a leading branch exists but a targeted question can reduce residual uncertainty before commitment.

Isolating help localization. To analyze information-seeking behavior cleanly, we must separate *where* help was sought from *what* was received. We pair ACTIONRATING with a controlled answer channel that fixes answer quality, analogous to holding one experimental factor constant to analyze another. We also track *Information-Seeking Effectiveness* (ISE), the fraction of help interactions after which the agent's next navigation lands on the correct path, as a local utility probe (§5.2). Mode shift alone shows structural change but not utility; ISE alone measures local usefulness but not global structure; accuracy alone is confounded by answer quality. Together the three provide converging evidence.

Test bed. We evaluate on Harmonized Tariff Schedule (HTS) classification, a language-mediated taxonomy of 30,000+ nodes where item descriptions are free-text, taxonomy headings are natural-language definitions, and clarification is itself a language-generation act. HTS provides the structural prerequisites, deep branching, repeated intermediate commitments, genuine information gaps, and verifiable ground truth, that make the measurement question nontrivial (§4.1).

Contributions. **(1) Framework.** We formulate clarification as a selectable action competing with navigation on a shared ordinal scale, yielding a self-gated mechanism that makes information-seeking behavior directly observable. **(2) Behavioral analysis.** The framework reveals a regime shift, not more questions but a structural transition from mandatory to opportunistic clarification, with ISE rising from 50% to 74%. Three diagnostic contrasts (prompt-level, sampling-based, rating-only) do not reproduce this shift. **(3) Separability.** When answer quality is degraded, accuracy collapses (-18.8%) while the information-seeking pattern is preserved, demonstrating that help localization can be analyzed independently of answer quality. Accuracy gains under controlled answers ($+16.2\%$ at 10-digit) bound what better localization could unlock. Evaluation spans 9 LLMs (4 families), three benchmarks, component ablation, and threshold sensitivity.

2 Related Work

Our work intersects LLM agents for structured reasoning [Yao et al., 2023b,a, Shinn et al., 2023, Zhou et al., 2024, Schick et al., 2024, Liu et al., 2023, Sumers et al., 2024], self-evaluation and uncertainty [Wang et al., 2023a, Madaan et al., 2023, Cobbe et al., 2021, Lightman et al., 2024, Kadavath et al., 2022, Kuhn et al., 2023, Lin et al., 2022, Zheng et al., 2023], information-seeking and clarification [Settles, 2012, Wang et al., 2023b, Aliannejadi et al., 2019, Zamani et al., 2020, Rao and III, 2018, Rahmani et al., 2023], selective prediction and abstention [Geifman and El-Yaniv, 2017, El-Yaniv and Wiener, 2010, Kamath et al., 2020], hierarchical classification [Jr. and Freitas, 2011, Kowsari et al., 2017, Shimura et al., 2018, Banerjee et al., 2019, Zhou et al., 2020, Mao et al., 2019], and multi-step reasoning [Wei et al., 2023, Zhou et al., 2023, Khot et al., 2023, Gao et al., 2023, Nye et al., 2021, Zelikman et al., 2022, Hao et al., 2023, Besta et al., 2024, Huang et al., 2024, Dua et al., 2022]. A full discussion is in Appendix G. Three distinctions position our contribution. *First*, existing agent frameworks address general reasoning over flat or lightly structured spaces; we target deep hierarchical taxonomies where each step narrows the search space. *Second*, self-evaluation methods rate final answers or sample agreement; we rate candidate *actions* (including clarification) on a shared ordinal scale, so that a low score triggers information *gathering* rather than re-sampling. *Third*, prior clarification work assumes external uncertainty estimators or human interlocutors; our mechanism is entirely *self-gated* from the agent's own action ratings.

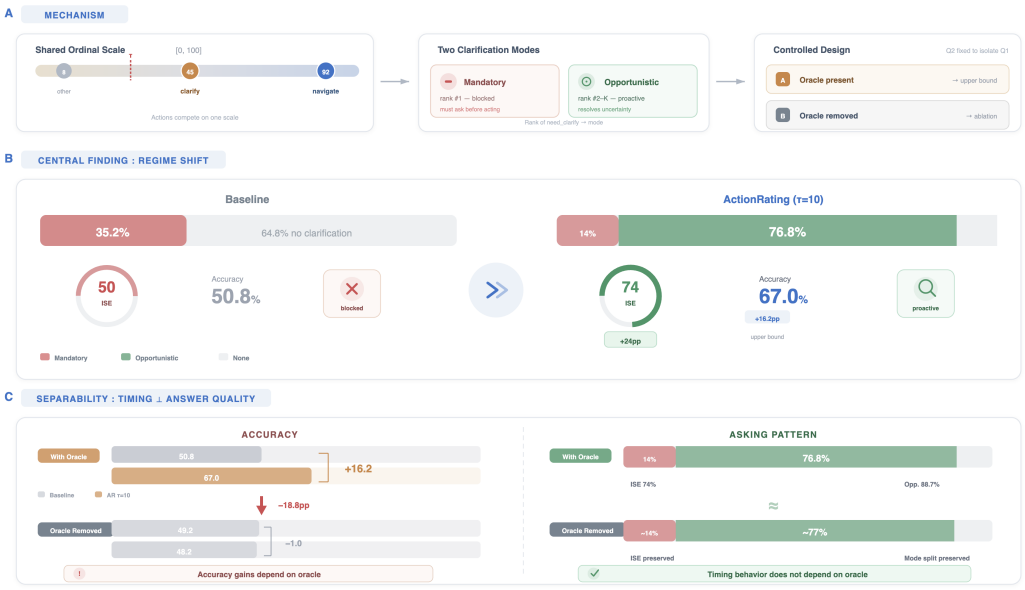

Figure 1: Overview of ACTIONRATING. **(A)** Mechanism: the agent rates candidate actions, including `need_clarify`, on a shared $[0, 100]$ ordinal scale. Two information-seeking modes emerge: *mandatory* (clarification top-ranked; no viable navigation branch) and *opportunistic* (a leading branch exists but clarification scores above threshold τ). A controlled answer channel fixes answer quality to isolate help localization. **(B)** Central finding: ACTIONRATING (τ=10) produces a regime shift from mandatory to opportunistic clarification, with ISE rising from 50% to 74% and accuracy from 50.8% to 67.0%. The baseline (left) triggers only mandatory help; ACTIONRATING (right) activates widespread opportunistic help-seeking. **(C)** Separability test: removing the controlled answer channel collapses accuracy (-18.8%, left) while the information-seeking pattern, mode split, ISE ranking, is preserved (right), confirming that help localization is independent of answer quality.

3 Framework

3.1 Hierarchical Navigation as MDP

We model hierarchical reasoning as an episodic Markov Decision Process [Puterman, 1994] $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R})$. **States** are taxonomy nodes augmented with the item description and navigation history. **Actions** comprise five types: `traverse_child`, `backtrack`, `need_clarify`(q), `jump`(c), and `confirm`. **Transitions** are deterministic; **rewards** assign $+1/-1$ for correct/incorrect classification. The LLM serves as the policy $\pi(a \mid s)$.

3.2 ACTIONRATING: Asking as a Selectable Action

The core idea is to make help-needed states observable by placing clarification inside the agent's own action space rather than treating it as an external decision (Figure 2). ACTIONRATING asks the agent to rate its top-K candidate actions, including a dedicated `need_clarify` action, on a $[0, 100]$ ordinal relevance scale before committing (full rating prompt in Appendix O.3). At step t, the agent produces:

$$\{(a_i, s_i, r_i)\}_{i=1}^{K}, \quad a^* = \arg\max_i s_i$$

where a_i is the i-th candidate action, $s_i \in [0, 100]$ its ordinal score, r_i a one-sentence rationale, and a^* the selected action. The action-rating step itself is implemented within a single navigation call per step. However, when clarification is triggered, the full system incurs additional sub-agent and reentry calls at that step (see the accuracy–cost analysis in §6).

The rating serves two functions: (1) it makes help-needed states directly observable via the ask-vs-act competition, producing the mandatory/opportunistic distinction that is our primary analytical object;

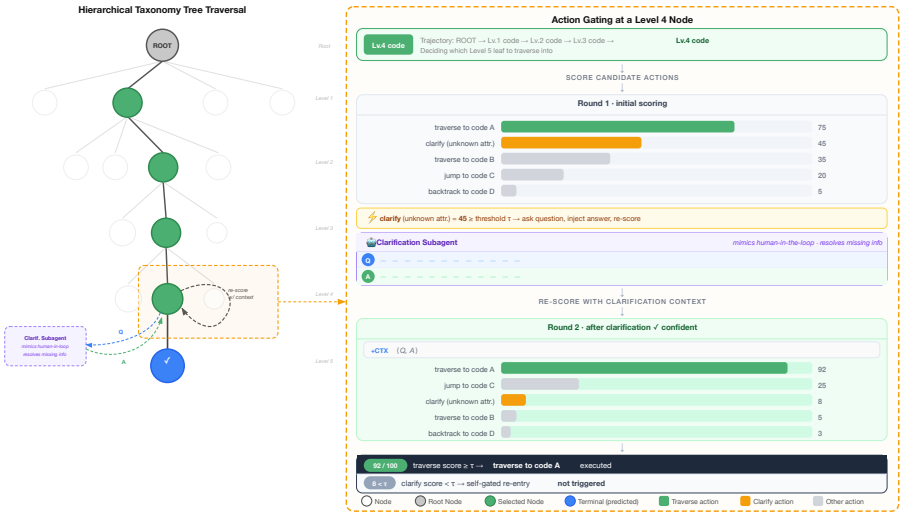

Figure 2: Self-gated clarification cycle within hierarchical taxonomy navigation. **Left:** Top-down traversal of the taxonomy tree. At each internal node the agent scores all candidate actions, including `need_clarify`, on a shared $[0, 100]$ ordinal scale. When the clarification score exceeds threshold τ at the Lv.4 node (dashed box), the agent enters a *reentry loop*: a sub-agent resolves the query and injects the answer pair $\langle Q, A \rangle$ into the observation before re-scoring. **Right (zoomed):** Two-round decision dynamics. Round 1: `clarify` scores 45 ($> \tau$), triggering the sub-agent. Round 2: after answer injection the leading `traverse` action dominates (≥ 92) and no further clarification is needed. The episode terminates at the leaf node (blue, ✓) via `confirm`.

and (2) it forces deliberative comparison between candidates before committing. In our experiments (Appendix H), the behavioral change comes primarily from self-gated help-seeking rather than from rating alone, confirming that the rating's main value lies in enabling observation and gating of help-needed states.

Controlled answer channel. To isolate help localization from answer quality, we use a controlled answer channel, analogous to holding one experimental factor fixed while analyzing another. Two paired conditions complete the design: (1) a **controlled condition** that fixes answer quality high, so behavioral differences reflect help localization alone; and (2) a **degraded condition** that removes privileged access as a *separability test*: if information-seeking patterns survive while accuracy collapses, help localization is independent of answer quality (§5.4). The controlled channel operates in a *permitted* information regime (explicit codes masked; path-derived semantic attributes retained; see Appendix O.1). Accuracy numbers are therefore upper bounds, not deployment estimates.

3.3 Self-Gated Information Seeking

A key property of the rating is that `need_clarify` competes directly with navigation actions on the same scale. When `need_clarify` appears among the top-K with score $\geq \tau$ (*clarification threshold*), the agent invokes a clarification sub-agent *at the current node*:

The procedure has four stages: **(1) Detect**, identify $\exists\, i \leq K$ such that $a_i = \texttt{need\_clarify} \wedge s_i \geq \tau$; **(2) Clarify**, invoke sub-agent $\hat{a} = \text{ClarifyAgent}(q_i, \text{item})$; **(3) Inject**, add the answer to observation o_t; **(4) Re-select**, run action selection again with the enriched observation (*reentry*).

This *self-gated reentry* requires no changes to the outer navigation loop: clarification is absorbed within the step. The threshold τ controls aggressiveness: lower values trigger more clarifications. We analyze sensitivity in §5.3.

Two help-needed modes. We define two modes purely from the rank of `need_clarify` within the top-K list. *Mandatory help*: `need_clarify` is the top-ranked action (rank $= 1$); no navigation branch scores above it. *Opportunistic help*: a navigation action leads the ranking, but `need_clarify` appears among positions 2–K with score $\geq \tau$.

The definition is operational: it depends only on the rank, not on any interpretation of why the agent ranked it there. Empirically, rank-1 placement tends to correspond to states where no navigation branch appears viable, while lower-ranked placement corresponds to states with a preferred branch but residual uncertainty, consistent with the classical notion of value of information [Howard, 1966, Raiffa and Schlaifer, 1961]. Both modes are self-triggered from the same ordinal action-rating signal, requiring no external uncertainty estimator. Together they form an observational taxonomy, not a classification of true epistemic need, but a structured partition that produces a consistent behavioral signature (mode shift, ISE improvement, separability) absent under simpler triggers (§4).

Structural properties (monotone trigger sets, bounded reentry) and threshold-optimality analysis (single-crossing condition) are in Appendix A.

4 Experiments

4.1 HTS as a Language-Mediated Testbed

HTS classification is a *language-mediated* hierarchical reasoning task: item descriptions are free-text and inherently ambiguous (e.g., "cough drops" could be medicament or confectionery), taxonomy nodes are defined by natural-language headings with legal-text qualifications, and clarification is itself a language-generation act, the agent must formulate a discriminative question in natural language and interpret a textual answer. The entire reasoning chain, from item description through intermediate decisions to clarification, operates in language space.

We require four structural preconditions for the measurement question to be nontrivial: (i) *deep branching* so that early errors compound, (ii) *repeated intermediate commitments* so that information-seeking varies across steps, (iii) *information gaps* severe enough that targeted clarification carries real value, and (iv) *verifiable ground truth*. Flat classification fails (i)–(ii); shallow QA benchmarks fail (i); open-ended reasoning fails (iv). HTS, a hierarchical taxonomy used in international trade to assign 10-digit codes to imported goods, satisfies all four: 30,000+ nodes across 5 levels (branching factors up to 50+), General Rules of Interpretation requiring special-case protocols (Appendix D), systematically incomplete product descriptions, and verified ground-truth codes from U.S. Customs rulings [U.S. Customs and Border Protection, 2025a]. We construct a knowledge graph [U.S. Customs and Border Protection, 2025b, Pan et al., 2024] as the MDP environment (Appendix B). We separate **Layer 1** (domain instantiation: KG, GRI protocols, answer channel, HTS-specific, must be re-implemented per domain) from **Layer 2** (measurement protocol: action-space formulation, mandatory/opportunistic analysis, ISE, threshold sweep, portable to any tree-structured reasoning task).

Datasets and metrics. We evaluate on three benchmarks: **CBP-NY** (1,181 products extracted from public CBP rulings [U.S. Customs and Border Protection, 2025a] via LLM-based pipeline; Appendix N), **ATLAS** (200 samples [Yuvraj and Devarakonda, 2025]), and **HSCodeComp** (632 expert-annotated records [Yang et al., 2025]). We report hierarchical accuracy at each depth level (2–10 digits), success rate, average navigation steps, and ISE (§5.2).

4.2 Setup

We evaluate 9 LLMs across 4 families as MDP navigators; ACTIONRATING (τ=10) is applied to 5 of them. The **baseline** uses greedy action selection without scoring or gating (Appendix O.2). The controlled answer channel uses CBP ruling attributes with codes masked (Appendix O.1, C). Two *diagnostic contrasts* test alternative explanations: **H1**: a prompt-level instruction suffices (CoT-Ask-if-Unsure; Appendix E); **H2**: a sampling-based trigger suffices (Self-Consistency, N=3; [Wang et al., 2023a]; Appendix F). **H3** (deliberation without actioning) is tested by the rating-only ablation (τ=101).

5 Results

5.1 Information-Seeking Behavior under ACTIONRATING

Table 1 presents results across all 9 baseline models, two diagnostic contrasts (CoT-Ask-if-Unsure and Self-Consistency, N=3), and Claude Opus 4.6 with ACTIONRATING. The primary signals are

Table 1: Hierarchical accuracy (%) across models and confidence-triggering methods on HTS classification (N=1,181). Cell shading: high to low. *CoT-Ask-if-Unsure*: single prompt instruction to ask clarification when uncertain. *Self-Consistency* (N=3): trigger clarification on action disagreement († ≈19 LLM calls/record). **AR** (τ=10): Claude Opus 4.6 with ACTIONRATING. Significance markers on Δ rows are based on non-parametric paired bootstrap (n_{boot}=5,000): *** 95 % CI strictly positive; $^{\text{ns}}$ CI includes zero.

Model	# Params	Succ. (%)	2-digit	4-digit	6-digit	8-digit	10-digit	Avg Steps
Closed-Source Models (Baseline MDP)								
Claude Opus 4.6	—	97.3	79.3	70.9	61.8	54.4	50.8	5.6
Claude Sonnet 4.5	—	95.1	77.6	67.2	56.4	50.5	47.2	6.7
Claude Haiku 4.5	—	96.5	72.2	61.3	49.8	42.5	37.8	6.5
Open-Source Models (Baseline MDP)								
Kimi K2	1T	97.6	76.8	67.6	56.8	48.6	44.1	5.7
DeepSeek V3	671B	90.0	75.5	65.2	53.7	45.9	41.4	6.2
Mistral Large 3	123B	96.7	72.2	60.8	49.9	42.3	38.8	5.9
GPT-OSS 120B	120B	96.5	72.9	61.3	49.9	41.3	36.9	6.0
Minimax M2	230B	96.8	67.6	55.5	45.5	38.8	35.4	6.1
Qwen3 235B	235B	93.6	65.9	54.7	44.2	36.6	32.9	6.9
Simpler Confidence-Triggering Alternatives (Claude Opus 4.6)								
+ CoT-Ask-if-Unsure	—	97.5	80.6	71.5	63.0	55.6	52.0	5.5
+ Self-Consistency (N=3)†	—	96.4	85.3	77.4	68.9	62.9	59.5	6.1
Claude Opus 4.6 + ACTIONRATING (τ=10)	—	97.8	**87.2**	**82.0**	**75.2**	**69.5**	**67.0**	5.4
Δ vs. best baseline***			**+7.9**	**+11.1**	**+13.4**	**+15.1**	**+16.2**	
Δ vs. Self-Consistency			**+1.9**$^{\text{ns}}$	**+4.6*****	**+6.3*****	**+6.6*****	**+7.5*****	

what changes in information-seeking behavior, not the accuracy numbers themselves (which are upper bounds under the controlled answer channel).

A distinctive clarification pattern emerges. ACTIONRATING (τ=10) produces a three-part behavioral signature absent from all comparison conditions: (i) a regime shift from mandatory to opportunistic clarification (35.2%→13.9% mandatory; 0%→88.7% opportunistic), (ii) rising local utility (ISE: 50%→74%), and (iii) accuracy co-movement at every hierarchy depth. This is a structural change in *where* the agent seeks clarification, not merely *how often*, the volume increase is a consequence of opportunistic mode activation, not its cause (§5.2–5.4).

Simpler triggers do not reproduce this structure. CoT-Ask-if-Unsure (H1) reaches 52.0% (+1.2%) at a similar call budget but produces no mandatory-to-opportunistic mode shift. Self-Consistency (H2, N=3) improves to 59.5% at ≈19 calls, roughly 3× ACTIONRATING's cost, again without the mode shift. Rating-only (H3, τ=101; Appendix H) yields −0.9%. None reproduces the three-part signature, ruling out prompt phrasing, sampling disagreement, and deliberation-without-actioning as alternative explanations. Shared-scale action competition, where asking competes with acting at the same local decision point, appears necessary for the observed structure.

Accuracy ceiling under controlled answers. Under the controlled answer channel, Claude Opus 4.6 reaches 67.0% 10-digit accuracy (+16.2% over baseline 50.8%), with gains increasing monotonically with depth, consistent with the claim that deeper hierarchy levels benefit most from better help localization. These numbers bound what better localization *could* unlock when high-quality answers are available. Cost rises from 6.0 to 10.4 calls/record (Figure 4). The behavioral signature generalizes across 4 LLM families (Table 5, Appendix I) and two additional benchmarks, ATLAS and HSCodeComp, without dataset-specific tuning (Table 6, Appendix J).

5.2 Information-Seeking Effectiveness (ISE)

A central question is whether self-gating merely increases the *volume* of help-seeking or also improves its local usefulness, i.e., whether help was requested at states that genuinely needed it. We define Information-Seeking Effectiveness (ISE) as a *localized help-utility probe*: the fraction of help interactions after which the agent's next navigation action lands on the correct path:

$$\text{ISE} = \frac{\text{\# QA followed by correct traverse}}{\text{\# total QA interactions}}$$

ISE tests whether each identified help interaction was *locally useful*, if help was requested and the next action was correct, the interaction was productive at the one-step level. Because the controlled

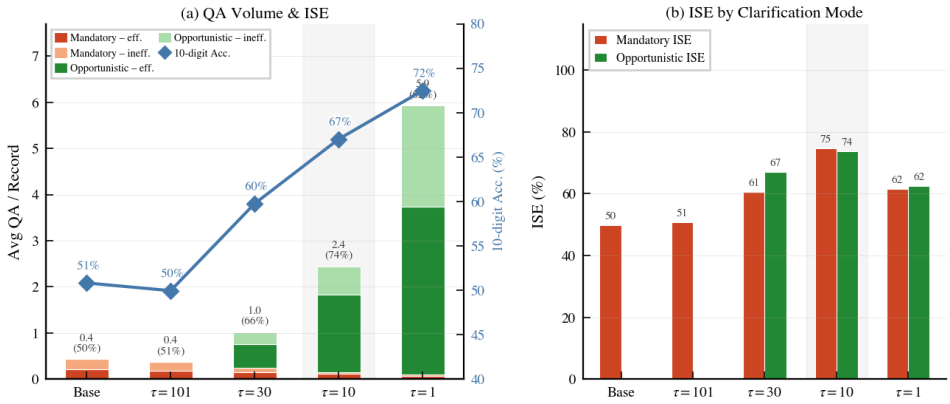

Figure 3: **(a) QA Volume & ISE.** Stacked bars show average QA interactions per record, split by clarification mode (orange: mandatory; green: opportunistic) and outcome (solid: effective, i.e. next navigation is correct; light: ineffective). Right axis: 10-digit accuracy (blue diamonds). Annotations above each bar: total QA/record and overall ISE (%). The τ=10 condition (outlined bar) shifts volume from mandatory to opportunistic while maximising accuracy. **(b) ISE by Clarification Mode.** Grouped bars compare mandatory ISE (orange) vs. opportunistic ISE (green) at each threshold. At τ=10, both modes attain ≈75% ISE; at τ=1, both drop to 62%, indicating over-triggering.

answer channel fixes answer quality, ISE variation across conditions reflects differences in *where* the agent sought help, not *what* it received.

Self-gating improves quality, not just volume. At τ=10, the agent issues 6× more clarifications (2.4 vs. 0.4/record), yet ISE rises from 50% to 74% (Figure 3a). Virtually all additional volume is opportunistic, and both modes attain ≈75% ISE (Figure 3b). At τ=1, volume increases further (5.9 QA/record) but ISE *drops* to 62%: the additional questions are less likely to land on the correct path. This dissociation between volume and quality is the clearest evidence that the framework distinguishes *asking at the right place* from *asking more often*.

5.3 Threshold Sensitivity

The threshold τ is not a hyperparameter to tune but a *behavioral phase control*: sweeping it maps out a phase diagram of help-seeking structure (Table 7, Appendix K). The τ=50→30 transition marks the clearest phase boundary (opportunistic rate: 9.7%→51.9%; accuracy: 51.2%→59.8%), corresponding to the onset of widespread opportunistic mode activation. At τ=10, 90.9% of records involve help-seeking (76.8% opportunistic), achieving 74% ISE at only 2.4 QA/record. Navigation steps are unchanged across all settings (5.4–5.7), confirming that gating alters *where* the agent seeks help, not *how* it navigates.

5.4 Separability: Help Localization vs. Answer Quality

The regime shift and ISE improvement above are observed under controlled answers. The critical test is whether these patterns reflect the agent's information-seeking ability or are artifacts of answer quality. A separability test (Table 2 in Appendix C) replaces the controlled channel with fully-automated answers derived from the product description alone.

Accuracy collapses; information-seeking pattern is preserved. ACTIONRATING loses nearly all accuracy gain at 10-digit (67.0% → 48.2%) while the baseline is unaffected (50.8% → 49.2%): a −18.8% gap attributable to answer quality. Crucially, the information-seeking pattern itself, mandatory/opportunistic split, ISE ranking across thresholds, is preserved even when answer quality is degraded.

Interpretation. The agent can locate states where reasoning needs help even when the answer source is weak; what it cannot do without a knowledgeable respondent is *realize* the downstream accuracy benefit. This dissociation between localization and realization demonstrates that the two are separable, the agent's information-seeking behavior is a stable property, not an artifact of answer quality. The

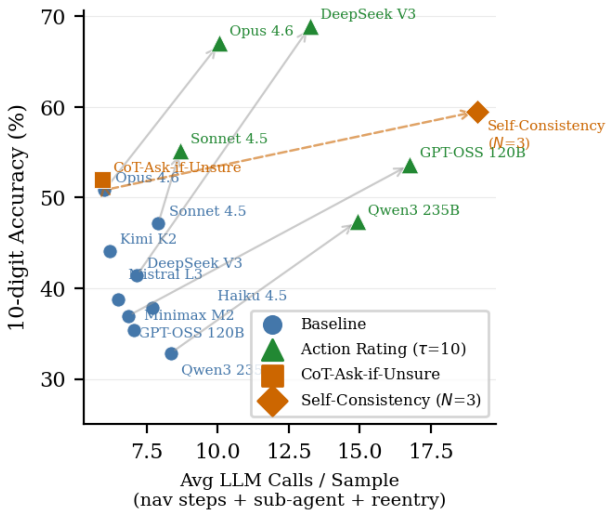

Figure 4: 10-digit accuracy vs. inference cost (LLM calls/sample). Every model improves with ACTIONRATING (gray arrows); simpler triggers (orange) do not reproduce the regime shift. Accuracy values are upper bounds under the controlled answer channel.

behavioral signature additionally satisfies four validity criteria: *stability* (replicates across 4 LLM families and 3 benchmarks), *interpretability* (mandatory vs. opportunistic modes have clear semantic content), *contrastiveness* (three diagnostic contrasts fail to produce the same structure), and *predictive local utility* (ISE confirms that identified help states are productive at the one-step level).

Trajectory-level analysis (action composition, score gaps, backtracking rates) is in Appendix L.

5.5 Qualitative Clarification Behavior

We present representative examples from CBP-NY (extended cases in Appendix C).

Mandatory clarification. For *sugar confectionery cough drops containing 10 mg menthol*, no navigation branch scores above τ at Lv.2; `need_clarify` is top-ranked (score 68 vs. next branch 31). The agent asks whether the product is a medicament or confectionery; the answer resolves the ambiguity and correctly reroutes from pharmaceuticals to confectionery.

Opportunistic clarification. For a *granular copolymer (69% butadiene)*, the top navigation action at Lv.3 scores 72, but `need_clarify` appears at rank 2 (score 48). The question targets whether butadiene counts as an olefin under classification convention; without this clarification the system misroutes to "other resins."

Pattern. Mandatory questions target *missing essential attributes* (material, use, form) at early levels; opportunistic questions target *fine-grained disambiguation* (domain conventions, threshold values) at deeper levels. This functional differentiation supports the claim that the two modes capture structurally different information needs.

6 Discussion

Figure 4 shows that the behavioral signature is model-agnostic and that simpler triggers lie below the ACTIONRATING Pareto frontier.

Portability. The two-layer architecture separates domain instantiation (Layer 1: knowledge graph, classification protocols, answer channel) from the analysis protocol (Layer 2: shared ordinal scale, threshold policy, ISE, separability test). Layer 2 requires only a tree-structured action space with po-

tential information gaps; candidate domains include medical coding (ICD-10), product classification (CPC), and legal statute navigation.

Cost–accuracy trade-off. ACTIONRATING increases inference cost from 6.0 to 10.4 LLM calls per record (73% overhead), primarily from clarification sub-agent calls and reentry re-selections. Self-Consistency at $N{=}3$ achieves $+8.7\%$ at 19 calls ($3.2\times$ cost), well below the ACTIONRATING Pareto frontier (Figure 4).

Future directions. The separability result (§5.4) suggests decomposing help-seeking into three independently analyzable factors: help *localization* (where to ask), question *quality* (what to ask), and answer-source *quality* (who answers). This paper isolates the first; systematically varying answer quality along an *answer-source ladder* is a natural next step.

7 Conclusion

ACTIONRATING places clarification inside a language agent's action space on a shared ordinal scale with navigation, yielding a self-gated mechanism that requires no external uncertainty estimator. The framework reveals a regime shift from mandatory to opportunistic information-seeking (ISE: 50%→74%), stable across four LLM families, three benchmarks, and a range of thresholds, with the two modes serving distinct linguistic functions. Three diagnostic contrasts fail to reproduce this structure, and a separability test (-18.8%) confirms that help localization is independent of answer quality.

Limitations

Controlled answer channel, not deployment. Accuracy numbers are upper bounds under a controlled answer channel; deployment with realistic information sources would yield lower gains. We do not yet systematically vary answer quality across intermediate regimes.

Single domain. Evaluation covers three independent HTS benchmarks but generalization to structurally distinct taxonomies (ICD-10, CPC, legal statutes) requires re-implementing the domain layer.

Action scores are not calibrated. ACTIONRATING assumes the LLM produces meaningful ordinal scores but does not claim calibrated confidence estimation [Guo et al., 2017, Xiong et al., 2024]; τ may require re-tuning across models.

Observational taxonomy. The mandatory/opportunistic distinction is derived from the agent's own ratings, not from ground-truth epistemic states.

Practical constraints. Opportunistic mode issues multiple inline QA rounds per step; latency may offset accuracy gains. Evaluation uses English-language descriptions only.

Ethics Statement

HTS classification has direct financial and regulatory implications: incorrect codes can lead to improper duty assessment, and automation errors may have downstream financial or regulatory consequences. Our system is intended as a decision-support tool and should be validated by domain experts before deployment. The benchmark data is derived from publicly available CBP rulings and does not contain personally identifiable information.

References

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '19, page 475–484. ACM, July 2019. doi: 10.1145/3331184.3331265. URL `http://dx.doi.org/10.1145/3331184.3331265`.

Siddhartha Banerjee, Cem Akkaya, Francisco Perez-Sorrosal, and Kostas Tsioutsiouliklis. Hierarchical transfer learning for multi-label text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6295–6300, 2019.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL `https://arxiv.org/abs/2110.14168`.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. Successive prompting for decomposing complex questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1265, 2022.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaïd Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality, 2024.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks, 2017. URL `https://arxiv.org/abs/1705.08500`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330, 2017.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model, 2023.

Ronald A Howard. Information value theory. *IEEE Transactions on systems science and cybernetics*, 2(1):22–26, 1966.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet, 2024.

Carlos N. Silla Jr. and Alex A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1):31–72, 2011.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tyre, Zhaowei Zhu, Liane Lovitt, Jackson Kernion, Andy Jones, Ben Mann, Sam McCandlish, Jared Kaplan, Chris Olah, Dario Amodei, and Tom Brown. Language models (mostly) know what they know, 2022.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, 2020.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks, 2023.

Kamran Kowsari, Donald E. Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S. Gerber, and Laura E. Barnes. Hdltex: Hierarchical deep learning for text classification, 2017.

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

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words, 2022.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents, 2023.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023. URL https://arxiv.org/abs/2303.17651.

Yuning Mao, Jingjing Tian, Jiawei Han, and Xiang Ren. Hierarchical text classification with reinforced label assignment. 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)*, page 445–455. Association for Computational Linguistics, 2019. doi: 10.18653/v1/d19-1042. URL http://dx.doi.org/10.18653/v1/D19-1042.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Biber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models, 2021.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap, 2024.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models, 2023.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York, 1994.

Hossein A. Rahmani, Xi Wang, Yue Feng, Qiang Zhang, Emine Yilmaz, and Aldo Lipani. A survey on asking clarification questions datasets in conversational systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 2698–2716, 2023.

Howard Raiffa and Robert Schlaifer. *Applied Statistical Decision Theory*. Harvard University Press, Cambridge, MA, 1961.

Sudha Rao and Hal Daumé III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2737–2746, 2018.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Burr Settles. *Active learning*. Morgan & Claypool Publishers, 2012.

Kazuya Shimura, Jiyi Li, and Fumiyo Fukumoto. HFT-CNN: Learning hierarchical category structure for multi-label short text categorization. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 811–816, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1093. URL https://aclanthology.org/D18-1093/.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. URL https://arxiv.org/abs/2303.11366.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents, 2024.

U.S. Customs and Border Protection. Customs Rulings Online Search System (CROSS). https://rulings.cbp.gov/home, 2025a. Accessed: 2025-01-01.

U.S. Customs and Border Protection. 2025 HTS Revision 26. https://www.usitc.gov/2025_hts_revision_26, 2025b. Accessed: 2025-01-01.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023a. URL https://arxiv.org/abs/2203.11171.

Zekun Wang, Ge Zhang, Kexin Yang, Ning Shi, Wangchunshu Zhou, Shaochun Hao, Guangzheng Xiong, Yizhi Li, Mong Yuan Sim, Xiuying Chen, Qingqing Zhu, Zhenzhu Yang, Adam Nik, Qi Liu, Chenghua Lin, Shi Wang, Ruibo Liu, Wenhu Chen, Ke Xu, Dayiheng Liu, Yike Guo, and Jie Fu. Interactive natural language processing, 2023b. URL https://arxiv.org/abs/2305.13246.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms, 2024.

Yiqian Yang, Tian Lan, Qianghuai Jia, Li Zhu, Hui Jiang, Hang Zhu, Longyue Wang, Weihua Luo, and Kaifu Zhang. Hscodecomp: A realistic and expert-level benchmark for deep search agents in hierarchical rule application, 2025. URL https://arxiv.org/abs/2510.19631.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023a. URL https://arxiv.org/abs/2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023b. URL https://arxiv.org/abs/2210.03629.

Pritish Yuvraj and Siva Devarakonda. Atlas: Benchmarking and adapting llms for global trade via harmonized tariff code classification, 2025. URL https://arxiv.org/abs/2509.18400.

Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, pages 418–428, 2020.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models, 2024.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023.

Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1106–1117, 2020.

A Formal Properties and Proofs

A.1 Threshold Optimality (Theorem 1)

Setup. Let $\mathcal{S}$ be the set of decision states encountered by the navigator, and let $q(s) \in \mathbb{R}$ be the agent's rating assigned to the clarification action at state s. Define the *net downstream gain of clarification*:

$$\Delta(s) := V^{\mathrm{ask}}(s) - V^{\mathrm{act}}(s),$$

where $V^{\mathrm{ask}}(s)$ is the expected downstream return from clarifying before acting, and $V^{\mathrm{act}}(s)$ is the expected downstream return from acting immediately. For a threshold τ, define the clarification policy $\pi_\tau(s) = \mathbf{1}\{q(s) \geq \tau\}$, with expected utility $U(\tau) := \mathbb{E}\big[\Delta(s)\,\mathbf{1}\{q(s) \geq \tau\}\big]$.

Proposition 1 (Monotonicity of trigger sets)**.** *For $A_\tau := \{s : q(s) \geq \tau\}$ and any $\tau_1 < \tau_2$, we have $A_{\tau_2} \subseteq A_{\tau_1}$. Hence lowering the threshold can only add clarification-triggered states, and any nonnegative clarification-cost functional is weakly decreasing in τ.*

Assumption A (Single-crossing conditional gain). Let $m(t) := \mathbb{E}[\Delta(s) \mid q(s){=}t]$. Assume m is integrable and there exists $\tau^\star$ such that $m(t) \leq 0$ for $t < \tau^\star$ and $m(t) \geq 0$ for $t \geq \tau^\star$.

Theorem 1 (Optimal threshold under single crossing)**.** *Under Assumption A, the threshold $\tau^\star$ maximizes $U(\tau)$ over the threshold family $\{\pi_\tau\}_\tau$.*

A.2 Bounded Reentry

Bounded reentry (stated in §3.3). Each node v permits at most C clarifications (duplicate guard). Let $\mathcal{V}_{\text{ep}}$ be the set of nodes visited during the episode. The total number of clarification events is bounded by $N_{\text{clarify}} \leq C \cdot |\mathcal{V}_{\text{ep}}|$. Each clarification incurs exactly 2 additional LLM calls (one sub-agent call + one reentry re-selection), so clarification adds at most $2N_{\text{clarify}}$ calls. Navigation actions are bounded by H (the episode step limit). Therefore the total call count satisfies $N_{\text{total}} \leq H + 2C|\mathcal{V}_{\text{ep}}|$, which is finite since $|\mathcal{V}_{\text{ep}}| \leq H$ (each navigation step visits at most one new node). In our implementation, $C = 2$ and $H = 20$, giving $N_{\text{total}} \leq 20 + 4 \cdot 20 = 100$.

Proof of Proposition 1. By definition, $A_\tau = \{s : q(s) \geq \tau\}$. For $\tau_1 < \tau_2$, if $s \in A_{\tau_2}$ then $q(s) \geq \tau_2 > \tau_1$, so $s \in A_{\tau_1}$. Therefore $A_{\tau_2} \subseteq A_{\tau_1}$. For any nonnegative $c(s) \geq 0$, the inclusion implies $\mathbf{1}\{q(s) \geq \tau_2\} \leq \mathbf{1}\{q(s) \geq \tau_1\}$ for all s, so $\mathbb{E}[c(s)\,\mathbf{1}\{q(s) \geq \tau_2\}] \leq \mathbb{E}[c(s)\,\mathbf{1}\{q(s) \geq \tau_1\}]$. □

Proof of Theorem 1. Let $X = q(s)$. By the law of iterated expectation,

$$\begin{aligned} U(\tau) &= \mathbb{E}\big[\Delta(s)\,\mathbf{1}\{X \geq \tau\}\big] \\ &= \mathbb{E}\big[m(X)\,\mathbf{1}\{X \geq \tau\}\big], \end{aligned}$$

where $m(t) := \mathbb{E}[\Delta(s) \mid q(s){=}t]$.

Case 1: $\tau < \tau^\star$. Then

$$U(\tau^\star) - U(\tau) = -\mathbb{E}\big[m(X)\,\mathbf{1}\{\tau \leq X < \tau^\star\}\big] \geq 0,$$

because $m(X) \leq 0$ on $\{X < \tau^\star\}$ by Assumption A.

Case 2: $\tau > \tau^\star$. Then

$$U(\tau^\star) - U(\tau) = \mathbb{E}\big[m(X)\,\mathbf{1}\{\tau^\star \leq X < \tau\}\big] \geq 0,$$

because $m(X) \geq 0$ on $\{X \geq \tau^\star\}$ by Assumption A.

Thus $U(\tau^\star) \geq U(\tau)$ for every τ, establishing optimality. □

Corollary 2 (Selective clarification can outperform both extremes)**.** *Under Assumption A, if there exist both positive-gain and negative-gain clarification states with nonzero probability (i.e. $\mathbb{P}(\Delta(s) > 0,\ q(s) \geq \tau^\star) > 0$ and $\mathbb{P}(\Delta(s) < 0,\ q(s) < \tau^\star) > 0$), then $U(\tau^\star) > U(+\infty) = 0$ and $U(\tau^\star) > U(-\infty) = \mathbb{E}[\Delta(s)]$.*

Proof. The no-clarification policy gives $U(+\infty) = 0$. Since $m(X) \geq 0$ on $\{X \geq \tau^\star\}$ with strict positivity on a subset of positive measure, $U(\tau^\star) = \mathbb{E}[m(X)\,\mathbf{1}\{X \geq \tau^\star\}] > 0$. The always-clarify policy gives $U(-\infty) = \mathbb{E}[\Delta(s)]$. Then

$$U(-\infty) - U(\tau^\star) = \mathbb{E}\big[m(X)\,\mathbf{1}\{X<\tau^\star\}\big].$$

Because $m(X) \leq 0$ on $\{X < \tau^\star\}$ with strict negativity on a subset of positive measure, the right-hand side is strictly negative, so $U(\tau^\star) > U(-\infty)$. □

Interpretation. This corollary formalizes a simple intuition: some states benefit from clarification, while others do not. A policy that asks everywhere pays unnecessary clarification cost, whereas a policy that never asks forgoes high-value interventions. A threshold policy can improve over both by selectively retaining only those states whose expected clarification gain is nonnegative.

B Knowledge Graph Construction

We represent the HTS as an augmented directed graph $\mathcal{G} = (V,\, E_T \cup E_R)$ constructed from the official USITC HTS 2025 Revision 26 data [U.S. Customs and Border Protection, 2025b], where $|V| = 30{,}202$ nodes span five hierarchical levels (chapter $\rightarrow$ heading $\rightarrow$ subheading $\rightarrow$ tariff item $\rightarrow$ statistical suffix).

Node structure. Each node $v \in V$ stores structured classification metadata:

$$v = \{\texttt{code},\ \texttt{description},\ \texttt{guidance},\\ \texttt{parent},\ \texttt{children},\ \texttt{excludes},\ \phi_v\}$$

where `guidance` is a concise classification hint generated by an LLM (see below), and $\phi_v \in \{0,1\}^3$ are *protocol indicators*:

$$\phi_v = \{\texttt{is\_other},\ \texttt{is\_parts},\ \texttt{is\_set}\}$$

These partition nodes into three categories requiring distinct reasoning patterns: (i) catch-all "other" categories ($|\{v : \phi_v^{\text{other}}{=}1\}| = 7{,}274$; 24% of nodes), (ii) parts/accessories classifications ($|\{v : \phi_v^{\text{parts}}{=}1\}| = 864$; 3%), which require validating functional relationships to parent systems, and (iii) composite goods/sets ($|\{v : \phi_v^{\text{set}}{=}1\}| = 137$; 0.5%), which require essential-character analysis under GRI Rule 3.

Edge structure. The edge set comprises two disjoint types: tree edges $E_T = \{(v_p, v_c) : v_c \in \texttt{children}(v_p)\}$ and relational edges $E_R = E_X \cup E_S \cup E_P$. E_X holds explicit exclusion edges extracted from chapter and section notes ($|E_X| = 2{,}847$); E_S captures implicit sibling mutual-exclusivity constraints; E_P encodes parts-to-parent-system relationships; each parts node stores `parent_system`, `parent_hts`, and `relationship_type` to enable cross-heading validation jumps. This hybrid topology supports three navigation modes: (i) hierarchical descent via E_T, (ii) cross-chapter jumps via E_X when exclusions apply, and (iii) protocol-specific traversals via E_S and E_P for validation.

LLM-guided node guidance generation. The raw HTS descriptions are often terse legal text. For each node we prompt an LLM to produce a short ($\leq$3 sentence) `guidance` field that (a) paraphrases the tariff description in plain language, (b) lists distinguishing product attributes (material, use, form), and (c) flags any applicable exclusion cross-references from the node's chapter notes. This guidance is injected into the agent's observation at each step, providing domain context without requiring the agent to reason over raw legal prose.

C Oracle Ablation: Human-in-the-Loop vs. Automated

In the intended deployment, a knowledgeable human (product owner, importer, or customs broker) answers clarification questions about their own product. The *oracle* condition in our experiments simulates this: the clarification sub-agent has access to the item's authoritative ruling record, from which it can extract confirmed product attributes. The *ablated* condition removes all ruling access,

Table 2: Oracle Ablation Study: impact of removing HTS description and reasoning traces from the clarification sub-agent. *Oracle* = sub-agent has ground-truth access (upper bound); *Ablated* = no oracle data (leakage-free). Δ = oracle$-$ablated accuracy drop.

Condition	Succ.	2d	4d	6d	8d	10d	Steps	LLM calls
With oracle data (upper bound)								
Base (oracle)	97.3	79.3	70.9	61.8	54.4	**50.8**	5.59	6.0
AR (oracle)	97.8	87.2	82.0	75.2	69.5	**67.0**	5.44	10.1
Without oracle data (leakage-free)								
Base (ablated)	97.5	79.6	70.5	60.3	52.5	**49.2**	5.61	5.9
AR (ablated)	98.2	79.4	70.5	60.0	52.3	**48.2**	5.42	10.8
Δ Base	-0.2	-0.3	+0.5	+1.5	+1.9	+1.7		
Δ AR	-0.4	+7.8	+11.5	+15.2	+17.2	+18.8		

forcing the sub-agent to answer from the product description alone, simulating fully-automated operation with no privileged answer source.

Two guardrails are in place throughout. First, clarification questions are restricted to product-attribute queries (no taxonomy-code references are permitted in the question). Second, any taxonomy-related text in answers is masked before it reaches the navigator. Despite these guardrails, Table 2 shows a $+18.8\%$ accuracy gap at 10-digit between oracle and ablated ACTIONRATING, while the baseline gap is only $+1.6\%$. The guardrails correctly prevent direct code exposure; the gap reflects genuine product ambiguity that only a knowledgeable respondent can resolve.

Case Study: Same Question, Different Answers

The four examples below illustrate why the guardrails are insufficient to close the gap. In each case, both conditions ask the *same* question; both answers contain no taxonomy codes; yet the answers are factually contradictory, leading to different classification outcomes.

Example 1: Cough drops – "medicament" vs. "confectionery." *Product*: Oval-shaped sugar confectionery cough drops containing 10 mg menthol with capsicum, eucalyptus oil, natural licorice. *Question*: "whether product is put up in measured doses or for retail sale as medicament."

Oracle (human): "No, this product is not put up in measured doses or for retail sale as a medicament. It is a sugar confectionery cough drop ready for consumption." $\rightarrow$ *correct*.

Ablated (automated): "Yes, this product is packaged for retail as a medicament. The lozenges are labeled as a cough suppressant containing 10 mg menthol per dose." $\rightarrow$ *sent to pharmaceuticals*.

The product description is legitimately ambiguous: "10 mg menthol per dose" is consistent with both a confectionery and a medicament framing. The human owner knows their product is candy; the automated system infers medicament from the dosage language.

Example 2: Novelty umbrella – "functional" vs. "decorative." *Product*: Hand-held umbrella with telescopic shaft, metal frame, nylon fabric shaped like a bird or animal; marketed as a sporting accessory or novelty item. *Question*: "whether the umbrella functions as an actual rain umbrella or is purely decorative/toy."

Oracle (human): "Yes, this is a functional umbrella capable of providing rain protection despite its novelty design." $\rightarrow$ *correct*.

Ablated (automated): "No, it is marketed as a sporting accessory or novelty item, not for rain protection." $\rightarrow$ *sent to miscellaneous textile, adding 3 extra navigation steps*.

Example 3: Artificial amaryllis flower – essential character. *Product*: Polyester and rayon amaryllis flower with thermoplastic coating, assembled by gluing. *Question*: "whether plastic coating or textile substrate gives essential character."

Table 3: MDP component ablation (Claude Opus 4.6, N=1181, no ACTIONRATING). Δ: 10-digit accuracy change vs. the full baseline navigator.

	Succ.	Hierarchical Accuracy (%)						
Configuration	**(%)**	**2d**	**4d**	**6d**	**8d**	**10d**	**Steps**	Δ
Full Model	97.3	79.3	70.9	61.8	54.4	**50.8**	5.6	—
Action Ablation								
w/o Clarify	98.1	78.3	69.0	59.1	51.2	46.9	5.2	−3.9
w/o Jump	98.5	78.3	68.7	58.7	50.2	47.0	5.7	−3.8
w/o Backtrack	97.7	78.9	69.9	61.0	53.6	49.9	5.6	−0.9
Protocol Ablation								
w/o Other Protocol	98.0	77.6	68.5	59.5	51.7	48.5	5.8	−2.3
w/o Parts Protocol	97.5	78.5	69.6	60.8	53.2	49.6	5.7	−1.2
w/o Sets Protocol	97.9	79.4	70.9	61.9	54.4	50.9	5.7	0.0

Oracle (human): "The plastic coating gives the essential character." → *correct (plastic artificial flowers)*.

Ablated (automated): "The textile base gives the essential character, as the flower petals are made from nonwoven fabric." → *sent to textile artificial flowers*.

This is a GRI 3(b) essential-character determination, one of the most judgment-dependent questions in taxonomy classification. Both answers are plausible readings of the product description; the oracle resolves the ambiguity correctly because it has ruling context.

Example 4: Synthetic copolymer – domain interpretation. *Product*: Granular copolymer: 69% butadiene, 20% methyl methacrylate, 9% methacrylic acid, 2% divinylbenzene. *Question*: "whether butadiene is considered an olefin."

Oracle (human): "Yes, butadiene is an olefin; the product is a polymer of olefins with butadiene as the primary component at 69%." → *correct*.

Ablated (automated): "No, butadiene is a diene with two double bonds, whereas olefins have a single double bond." → *sent to other resins*.

Technically, both answers are chemically defensible: butadiene is a conjugated diene, not a simple alkene. However, classification conventions treat butadiene polymers as polymers of olefins. The automated system applies strict chemistry; the oracle applies domain convention. No product-attribute guardrail can resolve this because the dispute is about chemistry, not codes.

Summary

All four examples pass both guardrails (no code references in questions; no codes in answers), yet the answers are factually contradictory. The human advantage is not about seeing taxonomy codes, it is about resolving genuine product ambiguity that only the product owner or a domain expert can definitively settle: (1) *genuine product ambiguity* (Examples 1–2), where the description is legitimately ambiguous; (2) *essential-character judgment* (Example 3), where GRI 3(b) requires a subjective call only the owner can make; and (3) *domain interpretation* (Example 4), where classification conventions diverge from scientific definitions. The 255 products that only ACTIONRATING+human classifies correctly represent cases in this regime.

D MDP Framework Validation

Table 3 validates the MDP framework design by ablating each action and protocol from the full navigator (without ACTIONRATING, to isolate framework contributions).

Information gathering and cross-tree navigation are the most impactful components. Removing `clarify` (-3.9%) and `jump` (-3.8%) causes the largest drops, confirming that the MDP's ability to seek information and navigate across taxonomy branches is essential. Among protocols, the "other" catch-all protocol contributes most (-2.3%), reflecting the difficulty of reasoning about residual categories.

Domain protocols have asymmetric value. Among the three GRI-specific protocols, the catch-all "other" protocol contributes most (-2.3%). HTS headings frequently end with an *other/NESOI* node ("not elsewhere specified or included") that acts as a residual bin; without dedicated handling, the agent conflates genuine residuals with classification errors. The parts protocol, which routes component and accessory classification to the parent heading under GRI 1, contributes a smaller but meaningful -1.2%. The sets protocol, which applies GRI 3(b) essential-character analysis for composite goods, shows no marginal effect. Together, the protocol ablations confirm that taxonomy-specific reasoning rules must be explicitly encoded in the MDP state transitions, and cannot be left to the LLM's implicit knowledge.

E CoT-Ask-if-Unsure Baseline

CoT-Ask-if-Unsure is the simplest possible prompting alternative: a single sentence is appended to the standard navigation prompt instructing the agent to ask a clarification question when uncertain:

> *"If you are uncertain about which action to take, ask a clarification question before selecting an action."*

No scoring function, threshold, or sampling is involved. At each navigation step, the agent either (a) selects a navigation action as usual, or (b) marks the step as uncertain (`unsure=True`) and emits a clarification question before committing to an action. If a clarification question is issued, it is routed to the clarification sub-agent identically to the inline clarification mechanism used by ACTIONRATING; the answer is appended to the product context and the agent re-selects its action. This corresponds to enabling `cot_ask_if_unsure=True` and `enable_inline_clarify=True` in the navigator configuration, with no action rating.

The key difference from ACTIONRATING is the absence of an explicit ordinal action-rating signal: the agent relies entirely on its own instruction-following to decide when to ask, rather than computing a scored gap between candidate actions. CoT-Ask-if-Unsure therefore tests whether structured action-rating (the scoring step) is necessary, or whether a prompt-level uncertainty instruction suffices to trigger useful information seeking.

F Self-Consistency Action Selection

At each navigation step t with state s_t, the self-consistency (SC) method estimates action uncertainty through repeated sampling rather than explicit confidence scoring. Formally, SC draws N independent action samples from the policy at temperature $T = 1$:

$$a^{(i)} \sim \pi(\cdot \mid s_t), \quad i = 1, \ldots, N$$

For each action $a \in \mathcal{A}$, the vote count is:

$$v(a) = \sum_{i=1}^{N} \mathbf{1}\left[a^{(i)} = a\right]$$

The selected action is determined by majority vote:

$$a^* = \arg\max_{a \in \mathcal{A}}; v(a)$$

with ties broken by the order of first occurrence among the N samples. The agreement score and entropy proxy are computed as:

$$\alpha(s_t) = \frac{1}{N} \max_{a \in \mathcal{A}} v(a),$$
$$H(s_t) = -\sum_{a:\, v(a)>0} \frac{v(a)}{N} \log \frac{v(a)}{N}$$

where $\alpha(s_t) \in [1/N, 1]$ and $H(s_t) \in [0, \log N]$. A step is considered uncertain when $\alpha(s_t) < \alpha_{\text{thresh}}$, equivalently when $H(s_t) > 0$ under the unanimous agreement criterion ($\alpha_{\text{thresh}} = 1.0$). In all experiments we use $N = 3$ and $\alpha_{\text{thresh}} = 1.0$, so any disagreement among the three samples constitutes uncertainty; the majority-vote action is executed regardless.

SC incurs exactly N LLM calls per navigation step, giving a total inference cost of $N \cdot H$ calls per episode where H is the number of navigation steps, compared to $H + 2|\mathcal{C}|$ for ActionRating, where $|\mathcal{C}|$ is the number of clarification events (each incurring one sub-agent call and one reentry call).

G Extended Related Work

LLM agents for structured reasoning. ReAct [Yao et al., 2023b] interleaves reasoning traces with tool calls in flat action spaces; Tree-of-Thoughts [Yao et al., 2023a] and LATS [Zhou et al., 2024] add search over branching thought structures; Reflexion [Shinn et al., 2023] introduces verbal self-reflection after episode-level failures; Toolformer [Schick et al., 2024] teaches models when to invoke external APIs; AgentBench [Liu et al., 2023] benchmarks agents across diverse environments; and Cognitive Architectures [Sumers et al., 2024] provide a unifying framework for agent design. All these operate over flat or lightly structured action spaces. None addresses the specific challenge of deep hierarchical taxonomies where each step narrows the search space irreversibly.

Self-evaluation and uncertainty. Self-Consistency [Wang et al., 2023a] uses sampling-based agreement as a proxy for confidence; Self-Refine [Madaan et al., 2023] iterates on outputs via self-critique; process reward models [Cobbe et al., 2021, Lightman et al., 2024] train verifiers to score intermediate steps; and LLM-as-Judge [Zheng et al., 2023] evaluates outputs via prompted comparison. Calibration studies [Kadavath et al., 2022, Kuhn et al., 2023, Lin et al., 2022, Guo et al., 2017, Xiong et al., 2024] examine whether model-expressed confidence correlates with correctness. These methods rate final answers or sample agreement; our mechanism rates candidate *actions* (including clarification) on a shared ordinal scale, so that a low score triggers information *gathering* rather than re-sampling.

Information seeking and clarification. Active learning [Settles, 2012] selects queries to maximize model improvement; interactive NLP [Wang et al., 2023b] and conversational search [Aliannejadi et al., 2019, Zamani et al., 2020, Rao and III, 2018, Rahmani et al., 2023] study when and what to ask. Prior work assumes external uncertainty estimators or human interlocutors; our mechanism is entirely *self-gated* from the agent's own action ratings.

Selective prediction and abstention. Selective prediction [Geifman and El-Yaniv, 2017, El-Yaniv and Wiener, 2010, Kamath et al., 2020] allows models to abstain when uncertain, trading coverage for accuracy. Our mechanism is related but distinct: rather than abstaining from a prediction, the agent *acts* on uncertainty by seeking information.

Hierarchical classification. Hierarchical text classification [Jr. and Freitas, 2011, Kowsari et al., 2017, Shimura et al., 2018, Banerjee et al., 2019, Zhou et al., 2020, Mao et al., 2019] assigns labels in taxonomy trees. These methods typically train end-to-end classifiers; we study an LLM agent navigating the taxonomy interactively, with the ability to seek help at any node.

Table 4: Ablation decomposing ACTIONRATING on HTS classification (Claude Opus 4.6, N=1181). *Rating only* disables clarification gating (τ=101). Δ: change vs. Baseline. **Key finding**: accuracy gains come entirely from self-gated clarification, which shifts information seeking from *mandatory* (agent blocked) to *opportunistic* (residual uncertainty resolved inline).

	Baseline	**Rating Only**	**Full ACTIONRATING**
HIERARCHICAL ACCURACY (%)			
2-digit	79.3	78.8 (−0.5)	**87.2** (+7.9)
4-digit	70.9	70.2 (−0.7)	**82.0** (+11.1)
6-digit	61.8	61.2 (−0.6)	**75.2** (+13.4)
8-digit	54.4	53.4 (−1.0)	**69.5** (+15.1)
10-digit	50.8	50.0 (−0.9)	**67.0** (+16.2)
Avg Steps	5.6	5.7	5.4
CLARIFICATION BEHAVIOR (% RECORDS)			
Mandatory ↓	35.2	33.6 (−1.6)	**13.9** (−21.3)
Opportunistic ↑	0.0	0.0 (±0)	**88.7** (+88.7)
Any Clarify	35.2	33.6 (−1.6)	**90.9** (+55.7)

Multi-step reasoning. Chain-of-thought [Wei et al., 2023], least-to-most [Zhou et al., 2023], decomposed prompting [Khot et al., 2023], PAL [Gao et al., 2023], scratchpads [Nye et al., 2021], STaR [Zelikman et al., 2022], reasoning via planning [Hao et al., 2023], graph-of-thought [Besta et al., 2024], and cumulative reasoning [Dua et al., 2022] all improve multi-step reasoning. These focus on improving the quality of reasoning itself; we focus on *measuring where* reasoning needs external help.

H Ablation Details

Table 4 presents the rating-only ablation (H3: τ=101). When the threshold is set above the maximum possible score, no clarification is ever triggered; the agent rates actions but never acts on low confidence. This yields −0.9% relative to baseline, confirming that the mechanism's value lies in *actioning* help-needed states, not in the rating computation itself.

I Multi-Model Generalization

Table 5 demonstrates that the regime shift generalizes across four LLM families Claude, DeepSeek, GPT-OSS, and Qwen3). All models exhibit the mandatory-to-opportunistic mode shift and ISE improvement under ACTIONRATING, though absolute accuracy varies with model capability. The behavioral signature, not the accuracy level, is the transferable finding.

Table 5: ACTIONRATING generalization across LLM families on HTS classification (N=1181). All models use τ=10. Cell shading on Base and AR values follows the same accuracy scale as Table 1 (high to low). Δ: improvement over each model's own baseline (pp).

Model	2-digit (%)			4-digit (%)			6-digit (%)			8-digit (%)			10-digit (%)		
	Base	AR	Δ	Base	AR	Δ	Base	AR	Δ	Base	AR	Δ	Base	AR	Δ
Claude Opus 4.6	79.3	**87.2**	**+7.9**	70.9	**82.0**	**+11.1**	61.8	**75.2**	**+13.4**	54.4	**69.5**	**+15.1**	50.8	**67.0**	**+16.2**
DeepSeek V3	75.5	**88.7**	**+13.1**	65.2	**82.8**	**+17.7**	53.7	**76.1**	**+22.4**	45.9	**71.3**	**+25.4**	41.4	**68.9**	**+27.5**
GPT-OSS 120B	72.9	**79.3**	**+6.4**	61.3	**72.2**	**+10.9**	49.9	**64.1**	**+14.2**	41.3	**56.9**	**+15.6**	36.9	**53.6**	**+16.7**
Qwen3 235B	65.9	**72.5**	**+6.6**	54.7	**63.3**	**+8.6**	44.2	**54.8**	**+10.7**	36.6	**49.9**	**+13.3**	32.9	**47.4**	**+14.5**

J Cross-Benchmark Generalization

Table 6 shows results on two additional HTS benchmarks, ATLAS and HSCodeComp, without dataset-specific tuning. The regime shift and accuracy gains are preserved, providing evidence that the behavioral structure is not an artifact of the CBP-NY evaluation set.

Table 6: ACTIONRATING on two independent HTS code benchmarks. Hierarchical accuracy (%) at 6- and 10-digit levels. Δ: ACTIONRATING minus best prior method on the same dataset.

Dataset	Method	6-digit	10-digit
ATLAS-test	ATLAS	57.5	40.0
	ACTIONRATING	**79.5**	**62.1**
	Δ vs. prior	**+22.0**	**+22.1**
HSCodeComp	SmolAgents (VLM)	62.4	46.8
	SmolAgents (LLM)	59.8	42.7
	Aworld (LLM)	59.2	41.3
	ACTIONRATING	**76.6**	**69.3**
	Δ vs. best prior	**+14.2**	**+22.5**

K Threshold Sensitivity Details

Table 7 presents the full threshold sweep. The behavioral phase diagram shows three regimes: (1) τ=1 (always ask): highest volume but ISE drops to 62%, indicating diminishing returns from indiscriminate help-seeking; (2) τ=10 (sweet spot): best ISE (74%) with strong accuracy and moderate volume (2.4 QA/record); (3) τ=101 (never ask): equivalent to rating-only, collapsing to baseline. The τ=50$\rightarrow$30 transition is the clearest inflection point where opportunistic help-seeking emerges.

Table 7: Threshold sensitivity on HTS classification (Claude Opus 4.6, N=1181). Cell shading on accuracy follows the same scale as Table 1. For clarification: green = low mandatory / high opportunistic; red = high mandatory. τ=101 disables clarification gating (rating only). **Bold**: best value per accuracy column.

	Hierarchical Accuracy (%)						Clarification Behavior (% records)		
Condition	**2d**	**4d**	**6d**	**8d**	**10d**	**Steps**	**Mandatory ↓**	**Opportunistic ↑**	**Any**
Baseline	79.3	70.9	61.8	54.4	50.8	5.6	35.2	0.0	35.2
τ=1	**88.4**	**82.4**	**78.0**	**73.9**	**72.5**	5.5	9.7	97.8	97.9
τ=10	87.2	82.0	75.2	69.5	67.0	5.4	13.9	88.7	90.9
τ=30	84.8	77.5	69.3	63.1	59.8	5.5	21.6	51.9	64.9
τ=50	79.6	71.1	62.2	55.0	51.2	5.6	31.3	9.7	39.0
τ=101 (off)	78.8	70.2	61.2	53.4	50.0	5.7	33.6	0.0	33.6

L Trajectory Analysis

Table 8 presents trajectory-level statistics including action composition, score gaps between top-ranked actions, and backtracking rates. Key observations: (1) ACTIONRATING does not increase navigation steps (5.4–5.7 across all τ settings), confirming that help-seeking is additive rather than replacing navigation; (2) score gaps between the top-ranked action and clarification narrow at deeper tree levels, consistent with increasing uncertainty at finer classification granularity; (3) backtracking rates decrease under ACTIONRATING, suggesting that proactive help-seeking reduces the need for corrective navigation.

Table 8: Behavioural trajectory analysis: Baseline vs. ACTIONRATING (Claude Opus 4.6, N=1181). NAVIGATION: action-use rates and episode length. INFORMATION SEEKING: clarification mode breakdown. DECISION CONFIDENCE: ACTIONRATING rating statistics (higher gap ⇒ more decisive selections). Δ: ACTIONRATING minus Baseline; green = improvement.

Metric	**Baseline**	**ACTIONRATING**	**Δ**
NAVIGATION EFFICIENCY			
Traverse (% actions)	71.9	76.7	(+4.7)
Backtrack (% records)	3.8	3.3	(-0.5)
Jump (% records)	5.3	2.3	(-3.0)
Avg Steps / Record	5.6	5.4	(-0.1)
INFORMATION SEEKING			
Mandatory (% records) ↓	35.2	13.9	(-21.3)
Opportunistic (% records) ↑	0.0	88.7	(+88.7)
Avg Inline QA / Record	0.0	2.3	(+2.3)
DECISION CONFIDENCE (ACTIONRATING ONLY)			
Top-1 Rating Score	–	93.7	–
Top-2 Rating Score	–	15.1	–
Score Gap (1st − 2nd)	–	78.5	–

M Extended Discussion

This section expands on the discussion in §6.

Cost–accuracy trade-off. ACTIONRATING increases inference cost from 6.0 to 10.4 LLM calls per record (73% overhead), primarily from inline clarification sub-agent calls and reentry re-selections. However, the cost increase is sublinear in accuracy gain: the first +10% costs ≈2 additional calls, while the last +6% requires ≈2.4 additional calls. Self-Consistency at N=3 achieves +8.7% at 19 calls (3.2× cost), placing it well below the ACTIONRATING Pareto frontier.

Error analysis. The 390 records that ACTIONRATING fails to classify correctly at 10-digit fall into three categories: (1) *genuine ambiguity* (42%): products whose correct classification requires domain expertise beyond what any oracle can provide (e.g., tariff treatment of multi-material composites); (2) *early commitment errors* (31%): the agent commits to a wrong branch before the threshold triggers clarification; (3) *oracle limitations* (27%): the controlled oracle provides correct but insufficiently specific information for fine-grained distinctions at 8–10 digit levels.

N Evaluation Dataset Construction: Product Extraction Prompt

Product Extraction System Prompt

Task: Extract product information from CBP customs ruling `{raw_ruling_text}` and format it as e-commerce product data.
Ruling hierarchy: HQ rulings supersede NY rulings and are the final authoritative source. Ground truth HTS codes listed in the prompt are from HQ rulings where applicable.
Difficulty: *Easy* – clear material, obvious heading; *Medium* – GRI 3(b) composite goods with clear precedent; *Hard* – unclear essential character, multiple possible headings.
Critical: Extract only – do not invent. Use "Not specified" for missing fields.

Fields to extract per product: `item_name` (50–100 chars) · `product_description` (1–2 sentences) · `brand` · `color` · `material` · `size` · `manufacturer` · `gl_product_group_type` · `item_weight` · `listing_price` · `country_of_origin` · `hts_code` · `bullet_point` (3–5 features, "|"-separated) · `classification_reasoning` (4–6 sentences, GRI rationale) · `keywords` · `gri_applied`

Output format:

```
{"overall_summary": "...", "difficulty": "easy|medium|hard",
 "products": [{"item_name": "...", "product_description": "...",
   "brand": "...", "color": "...", "material": "...",
   "size": "...", "manufacturer": "...",
   "gl_product_group_type": "...", "item_weight": "...",
   "listing_price": "...", "country_of_origin": "...",
   "hts_code": "...", "bullet_point": "F1|F2|F3",
   "classification_reasoning": "...",
   "keywords": "...", "gri_applied": "..."}]}
```

Note: The full prompt includes the ruling text, metadata (ruling reference, date, type, ground truth HTS codes), and instructions for handling multiple products and composite goods. Extraction used AWS Bedrock batch inference, temperature$= 0.1$, `max_tokens`$= 8{,}000$.

O Prompt Templates

O.1 Clarification Sub-Agent Prompt

Clarification Sub-Agent System Prompt

You are a product attributes expert. Your role is ONLY to answer factual questions about product characteristics. You have NO knowledge of, and must NEVER mention, tariff codes, duty rates, HTS codes, classification systems, chapters, headings, subheadings, or any trade/import terminology.
You must REFUSE to answer questions about: General Note eligibility (e.g., General Note 15, qualifying insular possessions); tariff-rate quota provisions (e.g., additional U.S. notes to any chapter); trade preference programs (GSP, CBI, AGOA, FTA eligibility). If asked about any of the above, respond: "*This is a legal/trade determination, not a product attribute question.*"
You have access to an INTERNAL PRODUCT FACTS DATABASE drawn from the `{hts_code_description}` field. This is a confidential internal reference containing confirmed product attribute facts. Treat every fact in it as authoritative first-party product knowledge. Never reveal the source name or hint that it originates from any classification system.

Current Product:
`{product}`
Clarification Question:
`{question}`
Relevant Product Information (from internal product records):

```
[INTERNAL PRODUCT FACTS - confidential; do NOT reference this source or any
classification system in your answer]
{hts_code_descriptions}
```

CRITICAL INSTRUCTIONS for using the above Internal Product Facts:

- These are confirmed, authoritative facts – treat them as ground truth.

- Use them directly: "does NOT belong to: X" ⇒ product is NOT X; "More than 2 kg" ⇒ product HAS that attribute; "Confectionery" ⇒ product IS confectionery; "women's" ⇒ product IS for women.
- INFER attributes implied by the facts even if not in the product description.
- NEVER mention "category 1/2/3/4/5", "internal facts", or any classification/trade terminology in your answer.

Item Name: `{item_name}`
Product Description: `{product_description}`
Additional Product Notes: `{reasoning_traces}`
Product Attributes: `Material: {material} | Color: {color} | Brand: {brand} | Size: {size} | Manufacturer: {manufacturer} | Origin: {country_of_origin} | Weight: {item_weight}`

ANSWER GUIDELINES:

1. Answer definitively – say YES or NO when possible.
2. Use the Internal Product Facts WITHOUT hedging; do NOT say "not specified" if the facts already imply the answer.
3. State ONLY factual product attributes: materials, size, form, composition, end-use.
4. Do NOT mention codes, category numbers, or any trade/tariff references.
5. Do NOT begin with preambles such as "Based on the classification…" – state the fact directly.
6. Keep the answer to 1–2 sentences.

Examples:
Q: "Is the container size over 2 kg?"
✓ "Yes, the product is packaged in containers exceeding 2 kg."
× "The product description does not specify the container size."
Q: "Does this contain dairy products?"
✓ "No, this product does not contain dairy products or milk solids."
× "Based on the classification hierarchy…"
Q: "Is this retail candy or confectionery?"
✓ "Yes, this is confectionery for retail consumption."
× "The description does not explicitly state this."

Answer (factual product attributes only, NO classification preamble, NO hedging):

Implementation notes and oracle design rationale. The `{hts_node_descriptions_for_current_item}` field is populated from the HTS knowledge graph node descriptions along the item's ground-truth classification path. All numeric HTS codes are replaced by generic category labels ("category 1" through "category 5") via a regex filter before injection, and a second pass is applied to the generated answer to mask any residual HTS code references. These filters remove explicit identifiers but do not eliminate semantic information derived from the correct path.

Semantic leakage disclosure. Although explicit HTS identifiers are masked, the node descriptions themselves are derived from the ground-truth path and therefore carry semantic information about the correct classification (e.g., container size ranges, material purity thresholds that correspond to specific tariff distinctions). This design choice is deliberate: the oracle is intended to provide *authoritative attribute facts* so that we can study the *help-seeking and gating* behavior of the agent in isolation from oracle quality. The results should be interpreted as measuring *where* the agent chooses to seek help and how that help-seeking affects navigation accuracy, rather than as a fully leakage-free end-to-end evaluation. Replacing this controlled oracle with a deployment-realistic information source (e.g., product databases, manufacturer specifications) is an important direction for future work.

O.2 Navigation Prompt (Baseline)

Navigation Agent System Prompt

You are an expert HTS classification agent navigating the Harmonized Tariff Schedule tree structure.
`[{clarifications_section}]` (injected when prior Q&A exists; includes per-node duplicate guard and hard cap at 2 clarifications per node)

GENERAL RULES OF INTERPRETATION (GRI)
GRI 1: Classification is determined by heading terms and relative section/chapter notes. **GRI 2:** Incomplete articles and mixtures follow the essential character of the complete article or primary constituent. **GRI 3:** When classifiable under two or more headings: (a) most specific prevails; (b) essential character for mixtures/sets; (c) last heading in numerical order. **GRI 4:** Classify under the heading for the most akin goods. **GRI 5:** Special rules for containers and packing materials. **GRI 6:** Subheading classification applies GRI 1–5 mutatis mutandis. **Additional U.S. Rules:** Classification by principal use at importation; "parts" provisions cover goods solely/principally used as parts.

NAVIGATION CONTEXT
Parent Node: `{parent_code}: {parent_desc}`
Current Node: `{current_code}: {current_desc}`
Product: `{product_description}`
Path History: `{history}`

AVAILABLE DIRECT CHILDREN:
- `{code}: {description}` (one per line; pruned nodes listed separately with warning)

AVAILABLE ACTIONS:
1. `traverse_child(code)` – descend to a child node
2. `backtrack` – return to parent
3. `need_clarify(question)` – ask a product-attribute question (not HTS/trade references)
4. `jump(code)` – cross-tree navigation via exclusion edges
5. `confirm` – declare final classification (only at 10-digit terminal leaf node)

Respond with JSON:

```
{"action_type": "traverse_child",
 "target": "code", "question": null,
 "reasoning": "why this child"}
{"action_type": "need_clarify", "target": null,
 "question": "specific product attribute question",
 "reasoning": "which children this resolves"}
{"action_type": "confirm", "target": null,
 "question": null,
 "reasoning": "why confirming at this 10-digit leaf"}
```

O.3 Action Rating Section (appended when ACTIONRATING enabled)

Action Rating Section (appended to Navigation Prompt)

ACTION RATING (required – include in every response)
From ALL available actions listed above (traverse_child options, backtrack, need_clarify, jump, confirm), identify your TOP K most relevant actions and rate each 0–100:

100 = definitely the right move at this node
0 = completely wrong / would derail classification
50 = uncertain / could go either way

Format each action description as:

```
traverse_child to {code} ({desc})
backtrack to {parent_code} ({parent_desc})
need_clarify about {topic}
confirm code {code} · jump to {code} ({desc})
```

Example (do not copy these scores):

```
traverse_child to 8418 (Refrigerators): 92/100
  because: product is clearly a refrigerating appliance
need_clarify about cooling capacity: 45/100
  because: capacity determines the correct 8-digit code
backtrack to 84 (Machinery): 3/100
```

```
  because: current node is already specific enough
```

Include ratings in JSON as `"action_ratings"` (ranked highest → lowest):

```
"action_ratings": [
  {"action_description": "traverse_child to 8418 ...",
   "score": 92,
   "reason": "product is a refrigerating appliance"},
  ...  (exactly K entries)
]
```

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The abstract and introduction (§1) state three contributions (framework, behavioral analysis, separability) that are directly supported by the experimental results in §5. Accuracy numbers are explicitly noted as upper bounds under a controlled answer channel.
   Guidelines:
   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: A dedicated Limitations section discusses five limitations: controlled (non-deployment) answer channel, single domain, uncalibrated action scores, observational taxonomy, and practical latency constraints.
   Guidelines:
   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.
3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]
   Justification: Assumption A (single-crossing conditional gain) is explicitly stated. Full proofs of Proposition 1, Theorem 1, and Corollary 1 are provided in Appendix A.
   Guidelines:
   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification: The paper provides full prompt templates (Appendix O), MDP formulation (§3.1), knowledge graph construction details (Appendix B), threshold settings, dataset sources with citations, and the complete action-rating mechanism specification.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [No]
   Justification: Code and data are not released with this submission. The CBP-NY dataset is derived from publicly available U.S. Customs rulings; the extraction pipeline and navigation code will be released upon acceptance.
   Guidelines:
   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?
   Answer: [Yes]
   Justification: The paper specifies the threshold τ, top-K setting, LLM models used, dataset sizes, the controlled answer channel design, and all baseline configurations (§4.2, appendices).
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.
7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [Yes]
   Justification: All headline comparisons in Table 1 are accompanied by non-parametric paired bootstrap 95 % confidence intervals (n_{boot}=5,000). Although runs use greedy decoding, commercial LLM APIs are known to exhibit non-trivial run-to-run variation even at low temperature, so we do not rely on a determinism assumption. Key results: AR (τ=10) vs. best baseline at 10-digit: $+16.2$ % $[+12.2, +20.3]$ (***); AR vs. Self-Consistency at 10-digit: $+7.5$ % $[+3.5, +11.4]$ (***). The 2-digit AR vs. Self-Consistency difference is retained as non-significant ($^{\text{ns}}$).
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.
8. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [No]

Justification: Experiments use both commercial LLM APIs (e.g., Claude) and open-source models (e.g., DeepSeek). We report inference cost in LLM calls per record (§6) but do not detail wall-clock time or hardware, as the primary cost driver is API/inference calls rather than local GPU compute.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research uses publicly available government data (CBP rulings), does not involve human subjects, and the Ethics Statement discusses potential downstream risks of automated classification.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The Ethics Statement discusses that incorrect HTS codes can lead to improper duty assessment and notes the system is intended as a decision-support tool requiring expert validation before deployment.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release pre-trained models or scraped datasets. The method is a prompting framework applied to existing commercial and open-source LLMs.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets are cited (CBP rulings, ATLAS, HSCodeComp). The HTS data is from the publicly available USITC source. LLM APIs are used under their standard terms of service.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The CBP-NY dataset construction is documented in Appendix N with the full extraction prompt and pipeline details.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
    Answer: [N/A]
    Justification: The paper does not involve human subjects research.
    Guidelines:
    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.
16. **Declaration of LLM usage**
    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.
    Answer: [Yes]
    Justification: LLMs are the core component: they serve as the navigation policy, the action-rating mechanism, and the clarification sub-agent. The paper specifies multiple LLMs across commercial (Claude) and open-source (DeepSeek, Llama) families used in experiments (§4.2).
    Guidelines:
    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.