HSCODECOMP: A Realistic and Expert-level Agent Benchmark for Hierarchical Rule Application

Anonymous ACL submission

Abstract

Current agent benchmarks focus on open-domain web navigation or structured data utilization. However, they neglect a critical capability required for numerous domain-specific applications (e.g., legal, medical and e-commerce): hierarchical rule application, where agents must strictly adhere to expert-written rules with implicit logic and vague boundaries. To bridge this gap, we introduce HSCODECOMP, a realistic benchmark requiring agents to assign 10-digit Harmonized System Codes (HSCode) to commercial products based on official tariff classification rules and noisy product descriptions. Sourced from real-world large-scale e-commerce platforms, HSCODECOMP comprises 632 product entries spanning diverse categories, with ground-truth HSCodes rigorously annotated by domain experts. Evaluations across 23 state-of-the-art LLMs and agents reveal a huge performance gap: best agent achieves only 46.8% 10-digit accuracy, significantly lagging behind human experts at 95.0%. Crucially, detailed analysis demonstrates the challenges of hierarchical rule application: standard test-time scaling strategies fail to yield improvements and excessive reasoning steps degrade accuracy. Code and benchmark will be publicly released.

1 Introduction

Recent advancements have empowered Large Language Model (LLM)-based agents to handle complex tasks involving navigating web environments (Wei et al., 2025; Mialon et al., 2023b) or accessing structured databases (Hu et al., 2025a; Yu et al., 2025a). However, a critical gap persists in numerous professional domains (e.g., legal, medical and e-commerce), where decision-making is governed not by retrieved knowledge or internal knowledge of models, but by rigorous expert-written rules (Sadowski and Chudziak, 2025). This capability involves navigating expert-written rules containing vague boundaries and implicit logic—a challenge largely overlooked by current agent benchmarks (Mialon et al., 2023b; Wei et al., 2025).

To formalize this challenge, we categorize agentic knowledge into three levels of complexity (Figure 1, left): (1) *Level 1: Open-Domain Data.* Understanding and reasoning over massive web content. Established benchmarks include GAIA (Mialon et al., 2023b) and BrowseComp (Wei et al., 2025); (2) *Level 2: Structured Data.* Precise utilization of structured data like databases and knowledge graphs, as seen in domain-specific benchmarks like MedBrowseComp (Chen et al., 2025b) and FinSearchComp (Hu et al., 2025a); and (3) *Level 3: Hierarchical Rules.* This level requires not only domain-specific knowledge retrieval but also precise navigation through hierarchical expert-written rules. It presents unique challenges: (a) accurate decision-making when rules contain vague natural language descriptions (Sadowski and Chudziak, 2025); and (b) reasoning about logical dependencies among rules, such as exception clauses and cross-references (Guha et al., 2023) shown in Figure 10.

To rigorously evaluate this overlooked *Level 3* capability, we introduce HSCODECOMP, the first expert-level benchmark designed to evaluate agents on Hierarchical Rule Application. The task involves predicting the 10-digit HSCode for e-commerce products—the global standard for international trade governance (Grainger, 2024). Constructed from real-world data spanning diverse product entries, HSCODECOMP comprises 632 carefully curated product entries. To ensure evaluation authority, unlike crowdsourced datasets, all ground-truth labels are rigorously annotated and cross-validated by domain experts.

We evaluate 23 state-of-the-art agents and systems on HSCODECOMP, including state-of-the-art LLMs and agentic frameworks. The results expose a substantial disparity: As shown in the Figure 1

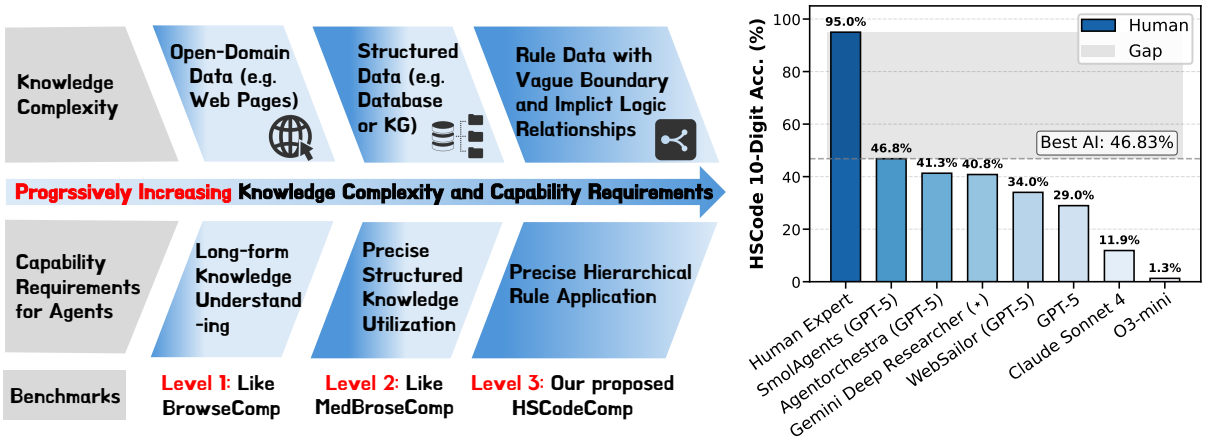


Figure 1. **Left:** Recent benchmarks reveal the increasing knowledge complexity and capability requirements for agents. **Right:** 10-digit HSCode accuracy of state-of-the-art baseline largely lags behind human experts (46.8% $<$ 95.0%), proving the challenges of hierarchical rule application. The closed-source agent ($\star$) is evaluated on the subset due to API unavailability.

(right), while human experts achieve 95.0% accuracy, the best-performing agent reaches only 46.8% (GPT-5). This massive gap highlights the inadequacy of current agents in handling hierarchical rule applications. Furthermore, detailed analysis uncovers several findings: standard test-time scaling strategies (e.g., majority voting (Liu et al., 2025; Zhu et al., 2025)) fail to improve performance, and excessive reasoning steps can degrade accuracy. These insights position HSCodeComp as a challenging testbed for current agents.

2 Related Works

HSCode Prediction. Most previous works treat HSCode prediction as the e-commerce text classification task (Grainger, 2024), using pre-trained BERT models (Liao et al., 2024; Shubham et al., 2022) or Large Language Models (LLMs) prompting (Gholamian et al., 2024). However, these approaches fail to leverage domain-specific knowledge, especially the rules written by human experts (Gholamian et al., 2024; Judy, 2024; Lee et al., 2024; Stassin et al., 2023).

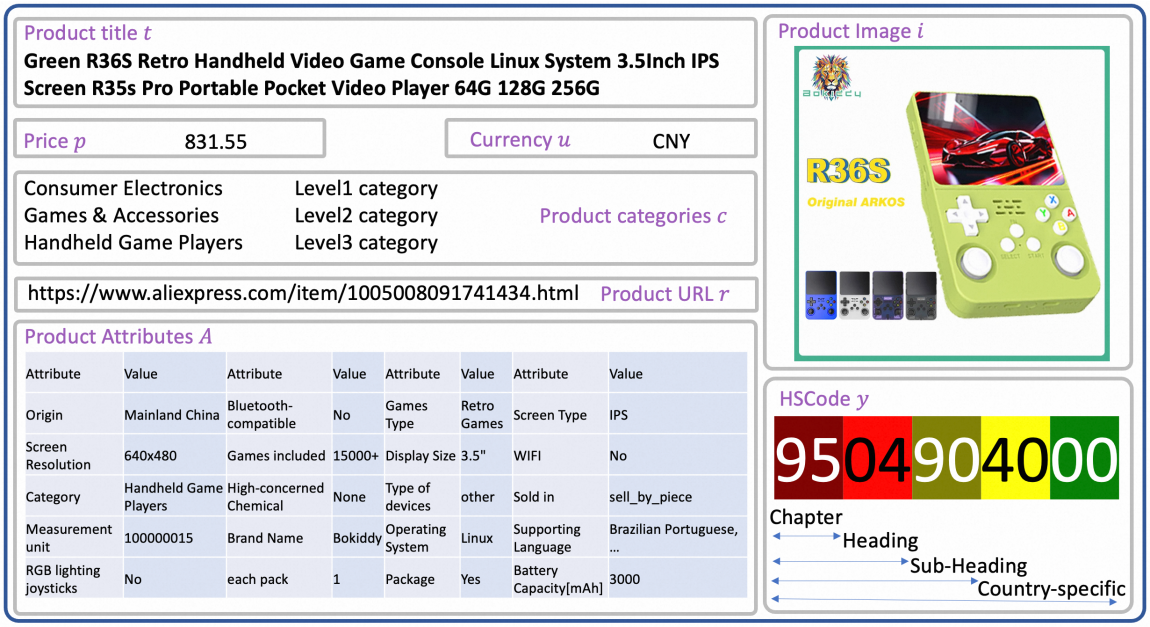
Benchmarking Level 1 Knowledge. Numerous benchmarks have been proposed to evaluate agent capabilities in understanding and deep reasoning over long-form open-domain web content (Thomas et al., 2025; Yao et al., 2024; Joshi et al., 2017; Phan et al., 2025). For example, WebArena (Zhou et al., 2023) provides realistic, self-hostable websites with standardized evaluation protocols to assess functional correctness. WebShop (Yao et al., 2022) and ALFWorld (Shridhar et al., 2021) evaluate long-horizon decision-making abilities of agents in web environments through tool interactions. More recent deep search benchmarks, such as GAIA (Mialon et al., 2023a), BrowseComp (Wei et al., 2025), WebWalkerQA (Wu et al., 2025) and BrowseComp-ZH (Zhou et al., 2025), demand advanced tool-usage and deep reasoning capabilities (Zhang et al., 2025; Li et al., 2025b).

Benchmarking Level 2 Knowledge. Recent works have focused on how agents utilize structured knowledge in domain-specific applications. Unlike open-domain data, domain-specific knowledge is typically organized into structured formats such as databases and knowledge graphs (Huang et al., 2025; Yu et al., 2025b; Chen et al., 2025a), enabling more precise knowledge retrieval and utilization. To evaluate these capabilities, numerous deep search benchmarks have been proposed, including WebMall (Peeters et al., 2025) and DeepShop (Lyu et al., 2025) for e-commerce, LegalAgentBench for law (Li et al., 2025a), FinSearchComp for finance (Hu et al., 2025a), and MedBrowseComp for medicine (Chen et al., 2025b).

In summary, while there exists numerous agent benchmarks that target open-domain or domain-specific scenarios, none evaluates hierarchical rule application capability. To address this, we introduce a realistic and expert-level benchmark HSCodeComp. Our benchmark presents significant challenges even for state-of-the-art closed-source and open-source agent systems.

3 Task Formulation

We formalize HSCode prediction as a knowledge-intensive hierarchical reasoning task. Unlike stan-



Attribute	Value	Attribute	Value	Attribute	Value	Attribute	Value
Origin	Mainland China	Bluetooth-compatible	No	Games Type	Retro Games	Screen Type	IPS
Screen Resolution	640x480	Games included	15000+	Display Size	3.5"	WIFI	No
Category	Handheld Game Players	High-concerned Chemical	None	Type of devices	other	Sold in	sell_by_piece
Measurement unit	100000015	Brand Name	Bokiddy	Operating System	Linux	Supporting Language	Brazilian Portuguese, ...
RGB lighting joysticks	No	each pack	1	Package	Yes	Battery Capacity[mAh]	3000

Figure 2. One example of a game console product in HSCODECOMP.

dard classification, this problem requires an agent not only to retrieve domain knowledge but also to strictly adhere to the expert-written hierarchical rules. Mathematically, the objective is to learn a mapping function $f : (\mathcal{X}, \mathcal{R}, \mathcal{K}) \to \mathcal{Y}$, where the agent predicts a valid HSCode $y \in \mathcal{Y}$ for product $x \in \mathcal{X}$ by applying hierarchical rules $\mathcal{R}$ supported by domain-specific knowledge $\mathcal{K}$.

Input Space ($\mathcal{X}$). As shown in Figure 2, Each input instance $x \in \mathcal{X}$ represents a commercial product profile containing real-world and noisy information. We define $x = (t, A, c, i, p, u)$, where: t is the product title; $A = \{(k_j, v_j)\}_{j=1}^{K}$ denotes K structured attributes (*e.g.,* material and package size); c represents product categories defined by the e-commerce platform; i is the product image; capturing visual features (e.g., texture, shape) absent in text; p and u are the price and currency.

Hierarchical Rules ($\mathcal{R}$). The core of HSCODECOMP is the Hierarchical Tariff Rules ($\mathcal{R}$), consisting of two kinds of rules: (1) **Tariff rules** are sourced from official eWTP platform, aggregating authoritative US tariff systems. HSCODECOMP evaluates agents on using US tariff rules. As shown in Figure 2, $\mathcal{R}$ is characterized by: (a) Hierarchical Structure: organized into Chapters (2-digit), Headings (4-digit), Sub-headings (6-digit) and Country-specific codes (10-digit); (b) Implicit Logic: frequently contain exclusion clauses and cross-references; and (3) Vague Boundaries: Linguistic descriptions often ambiguous definitions, requiring agents to interpret the boundary between categories; and (2) **Expert-written decision rules** provides the guidelines to resolve conflicts when multiple headings seemingly apply, which is a consolidated HSCode classification protocol distilled from years of accumulated domain expertise. Partial case can be found in Figure I.1.

Domain-Specific Knowledge ($\mathcal{K}$). To correctly interpret and apply $\mathcal{R}$, agents must access external domain knowledge ($\mathcal{K}$), i.e., a repository of historical customs rulings target US rulings, i.e., the CROSS system[1]. These rulings serve as few-shot examples, helping agents navigate edge cases where rules are ambiguous.

Output Space ($\mathcal{Y}$). The target $y \in \mathcal{Y}$ is a unique 10-digit numeric string. A valid prediction must represent a path in the HS taxonomy tree, satisfying all constraints defined in $\mathcal{R}$.

4 HSCODECOMP Benchmark

4.1 Benchmark Construction

We design a rigorous pipeline to ensure HSCODECOMP is diverse, realistic and expert-level. The construction process comprises three phases.

Realistic Sourcing and Filtering. We source product entries from a large-scale global e-commerce platform, adopting a sampling strategy anchored to actual sales volume and category statistics to reflect real-world distributions. Distinct

[1] https://rulings.cbp.gov/home

from synthetic datasets, we explicitly retain intrinsic noisy metadata (e.g., verbose titles, inconsistent attributes) to preserve the complexity of realistic scenarios. To ensure taxonomic diversity and prevent topical skew, we apply a semantic redundancy filter, discarding instances that share identical categories and HSCodes with existing entries. his ensures the benchmark captures the challenges of valid long-tail distributions rather than being dominated by frequent and trivial items.

Human Expert Annotation. We engage human experts specialized in HSCode classification to annotate the HSCode (y) for each product profile (x). As shown in Figure 3 (left), the annotation process follows a strict protocol: (1) Two experts gather comprehensive information from the product webpage (Step 1); (2) They extract the core structured features of products (Step 2); (3) Experts query the official ruling databases (CROSS) for related cases. If a related case is found, the corresponding HSCode is then validated on eWTP system, followed by the revision. Otherwise, they refine their search queries or revisit Step 2 to adjust the extracted features (Step 3); (4) For products without any related cases, experts execute expert-written decision rules to apply tariff rules, and determine the appropriate HSCode (Step 4); and (5) Finally, experts verify the final identified HSCodes on the eWTP website to ensure its validity (Step 5).

Human Expert Validation. To guarantee label reliability, we implement a dual-verification mechanism. As shown in the Figure 3 (Step 6), Two experts independently annotate each sample; if their codes match, the instance is accepted. In cases of disagreement, a senior tariff expert adjudicates the conflict. If no consensus is reached, the sample is discarded. Furthermore, to verify the reliability of our process, we conduct an additional quality review. A fourth senior expert, not involved in the initial annotation, re-annotated a random 10% sample of the dataset. This review shows only a 2% disagreement rate, confirming the effectiveness and consistency of our dataset construction pipeline.

Benchmark Statistics. The final dataset comprises 632 high-quality entries, spanning 27 HS chapters and 32 first-level product categories. Confidence intervals analysis in Appendix E proves the stable evaluation. More statistics of HSCODECOMP are shown in Appendix D.

Reproducibility. To ensure reproducibility amidst dynamic environments, we rigorously anchor all ground-truth annotations and knowledge bases to a fixed timestamp (June 30, 2025), aligning with community standards for temporal robustness (Wong et al., 2025).

4.2 Evaluation Metric

We adopt Exact Match Accuracy as the primary metric. Given the hierarchical nature of HS Codes, we report accuracy at multiple granularities: 2-digit (Chapter), 4-digit (Heading), 6-digit (Sub-heading), and the full 10-digit (Country-specific) level. The 10-digit accuracy serves as the strictest measure of an agent's ability to perform deep, end-to-end hierarchical reasoning.

5 Experimental Setup

We conduct a comprehensive evaluation across three distinct paradigms.

Foundation Models. We test 14 leading LLMs and VLMs (e.g., GPT-5, Gemini-2.5-Pro, Kimi-K2 (Team et al., 2025), Claude-Sonnet-4, DeepSeek series (DeepSeek-AI et al., 2025), Qwen series (Yang et al., 2025), o3-mini, Nemotron-32B (Bercovich et al., 2025) etc. relying solely on internal knowledge. For VLMs, product images are provided to assess visual reasoning capabilities.

Open-source Agent Frameworks. We evaluate six advanced frameworks (e.g., SmolAgents (Roucher et al., 2025), Aworld (Yu et al., 2025c), Agentorchestra (Zhang et al., 2025), OWL (Hu et al., 2025b), WebSailor (Li et al., 2025b) and Cognitive Kernel (Fang et al., 2025)) using GPT-5 as the default backbone. All frameworks utilize standardized tools including Google Search, Knowledge and Tariff Rules Retrieval.

Closed-source Agent Systems. Following previous works (Li et al., 2025b), we assess the performance of commercial systems (Manus, Gemini Deep Research, and Grok DeepSearch) on a representative subset of benchmark.

All systems are evaluated using the strict 10-digit Exact Match Accuracy metric. More implementation details appear in Appendix I.

6 Main Experimental Results

In this section, we present a rigorous evaluation of state-of-the-art systems on HSCODECOMP. We analyze performance across three dimensions: (1)

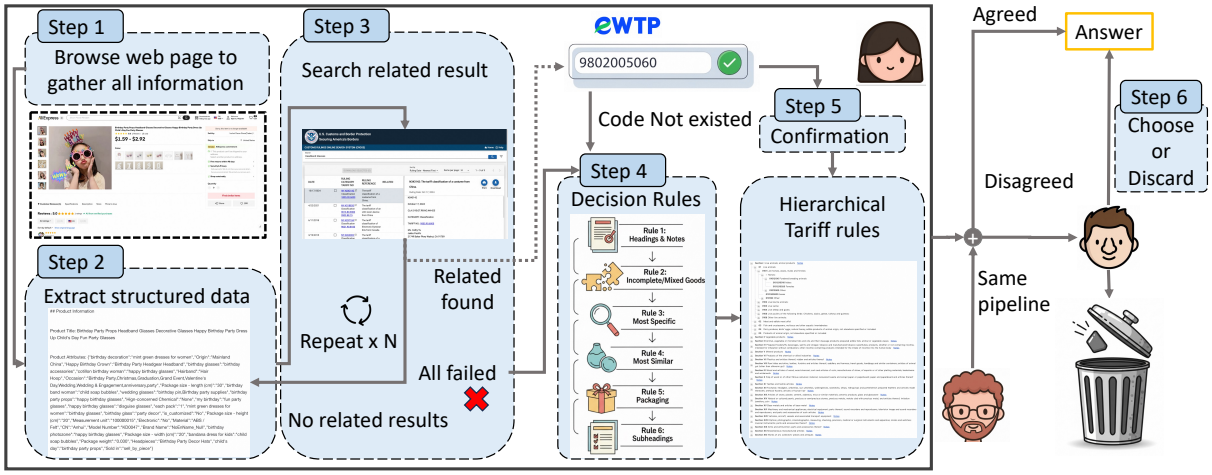


Figure 3. The pipeline for human experts to annotate the HSCodes, including two human experts for HSCode annotation (Step 1 to 5) and one additional expert for quality validation (Step 6).

the absolute gap between state-of-the-art agents and human experts; (2) the comparison analysis of open-source versus closed-source frameworks; and (3) backbone model selection.

6.1 Human-Agent Performance Gap.

Table 1 reveal a substantial disparity between agents and human experts in hierarchical rule application: (1) **Significant Performance Gap:** The state-of-the-art system, `SmolAgents` (GPT-5 w/ Vision), achieves only **46.83%** 10-digit accuracy, significantly lagging behind the **95.0%** accuracy of human experts, highlighting that current agents struggle to satisfy the rigorous constraints of expert-level tasks; (2) **Hierarchical Challenges:** Performance decays consistently as taxonomic depth increases. While agents maintain competitive accuracy at the 2-digit Chapter level (∼82%), they fail to navigate the fine-grained logical boundaries required for 10-digit country-specific classification. Moreover, to further accurately diagnose where the reasoning breaks, we further analyze the Conditional Accuracy. This metric measures the probability of correctly predicting a more fine-grained code (e.g., 4-digit) contingent upon the correctness of its preceding digits (e.g., 2-digit). As shown in Table 10, it can be found that current agents struggle with the complex, multi-step reasoning required for precise rule application at deeper hierarchical levels; and (3) **Agents Outperform LLMs:** Agentic frameworks consistently outperform standalone foundation models (e.g., SmolAgents 42.72% vs. GPT-5 LLM 28.96%), validating the necessity of tool usage like domain-specific knowledge retrieval and tariff rules retrieval.

6.2 Open-source vs. Closed-source Agents

We compare open-source frameworks against commercial "Deep Research" systems on a representative subset. Surprisingly, open-source agents demonstrate superior rule adherence. As detailed in Table 2, `SmolAgents` and `AWorld` (both ∼42.9%) outperform proprietary systems like Gemini Deep Research (40.8%) and Manus (30.6%). Qualitative analysis suggests commercial systems often suffer from premature commitment and information misprocessing, as detailed in Section 7.4.

System Architecture	10-digit Acc.
Closed-Source Commercial Systems	
Gemini Deep Researcher	40.81
Grok DeepSearch	26.53
Manus	30.61
Open-Source Frameworks (w/ GPT-5)	
SmolAgents	**42.86**
AWorld	**42.86**

Table 2. Performance comparison: Commercial Systems vs. Open-Source Frameworks.

6.3 Impact of Backbone Models

We evaluate the *SmolAgents* framework across four state-of-the-art backbone models to analyze the contribution of the underlying backbone models. Results in Table 3 demonstrate that GPT-5 is currently the only viable backbone for this high-complexity task. It achieves 42.72% accuracy, establishing a decisive lead over Gemini-2.5-Pro (34.49%), Claude 3.5 Sonnet (33.70%), and Qwen-Max (17.43%). Therefore, we choose GPT-5 as

Baselines	Model Type	HSCode Prediction Accuracy				
		2-digit	4-digit	6-digit	8-digit	10-digit
LLM/VLM-Only						
GPT-5	VLM	82.12	70.89	59.97	41.46	29.27
GPT-5	LLM	82.59	69.78	56.33	40.98	28.96
Gemini-2.5-PRO	VLM	**82.28**	71.04	59.02	40.51	24.21
Gemini-2.5-PRO	LLM	80.54	69.94	58.54	40.35	23.42
GPT-4o	VLM	78.01	64.08	48.10	29.75	18.51
GPT-4o	LLM	75.47	61.55	45.73	30.06	18.35
Claude Sonnet 4	VLM	78.80	64.08	45.25	22.63	11.23
Claude Sonnet 4	LLM	78.80	62.97	44.94	23.58	11.87
Kimi-K2	LLM	78.01	62.03	44.15	24.53	12.18
DeepSeek-R1	LLM	77.22	61.71	38.45	16.77	6.65
Qwen-Max	LLM	71.52	48.58	24.21	11.23	3.80
O3-mini	LLM	77.22	56.17	24.53	6.65	1.27
Open-source Agent System (GPT-5 Backbone)						
SmolAgents	VLM	82.06	**72.06**	**62.38**	**52.38**	**46.83**
SmolAgents	LLM	**82.28**	70.89	59.81	49.05	42.72
Aworld	LLM	**82.28**	70.41	59.18	48.58	41.30
Agentorchestra	LLM	82.12	70.73	60.44	47.78	41.30
OWL	LLM	72.63	61.87	51.58	41.77	37.34
WebSailor	LLM	81.64	70.56	57.27	43.98	35.44
Cognitive Kernel	LLM	80.06	69.15	54.59	40.03	26.42

Table 1. The complete results of state-of-the-art baselines in our proposed HSCODECOMP.

Backbone Model	10-digit (%)	Δ
GPT-5 (Default)	**42.72**	–
Gemini-2.5-Pro	34.49	−8.23
Claude 4 Sonnet	33.70	−9.02
Qwen-Max	17.43	−25.29

Table 3. Ablation on backbone models. Δ denotes the performance gap compared to GPT-5.

the default setup for open-source agents. Complete results are provided in Table 12 in Appendix G.3.

7 Analysis on HSCODECOMP

We conduct a multi-dimensional analysis characterizing both agent behaviors and benchmark properties. Our investigation reveals the limitations of current agent systems regarding hierarchical rule application: we find that standard test-time scaling strategies are ineffective, and excessive reasoning steps degrade performance due to reasoning drift. To identify the essential components for success, we conduct rigorous ablation studies that isolate and quantify the distinct roles of rules, knowledge and visual signal of products. Finally, through a comprehensive failure taxonomy and per-category analysis, we underscore the inherent complexity and challenging nature of HSCODECOMP.

7.1 Inefficacy of Test-Time Scaling

While test-time scaling (TTS) has proven effective for reasoning tasks (DeepSeek-AI et al., 2025), our experiments reveal that two standard TTS strategies fail to improve performance on HSCODECOMP (Liu et al., 2025): (1) **Majority Voting:** We implement majority voting across $K = \{1, 2, 4, 8, 16\}$ trials (Voting@K), using SmolAgent (GPT-5). Figure 4 showns that increasing K yields negligible improvement; and (2) **Self-Reflection:** We integrate a self-reflection mechanism into SmolAgent (GPT-5), enabling the model to proactively evaluate and revise its reasoning and actions. However, this approach slightly decreases performance from 42.72% to 42.57%. These results highlight the need of more effective test-time scaling strategy for hierarchical rule application.

7.2 Overthinking Decrease Performance

Current agentic paradigms often operate on the assumption that providing richer reasoning improves decision quality. As shown in Table 1, the reasoning-heavy WebSailor framework signif-

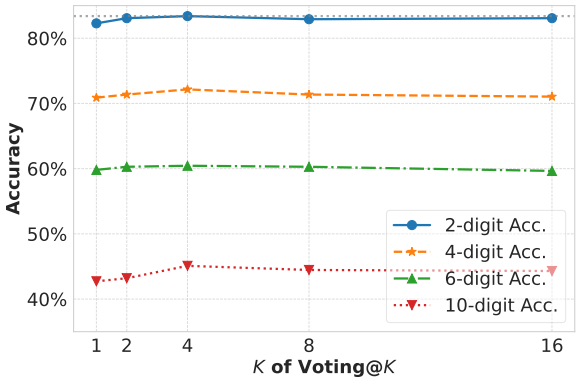


Figure 4. Test-time scaling (majority voting (Zhu et al., 2025)) analysis on HSCODECOMP.

icantly underperforms the direct-execution SmolAgents (35.44% vs. 42.72%). One case study in Appendix K suggests a phenomenon of "Reasoning Drift": without the grounding of accurate tool feedback (e.g. rules definitions), agents tend to hallucinate constraints or misinterpret vague product details. To validate this, we conducted an ablation study on WebSailor by constraining its reasoning depth: (1) **No-Think** (direct tool invocation); and (2) **Medium-Think** (moderate reasoning controlled by the prompt). As shown in Table 4, restricting reasoning depth yields significant gains, with the No-Think variant approach SmolAgents. This empirical evidence underscores a critical design principle: for domain-specific tasks where valuable information resides in domain knowledge and rules, prioritizing knowledge and rules utilization over internal reasoning is essential to prevent hallucination and reasoning drift.

Agent System	10-digit Accuracy (Depth →)			Gain
	Overthink	Medium	No-Think	
SmolAgents	–	–	**42.72**	–
WebSailor	35.44	37.34	**40.82**	+5.38

Table 4. Impact of reducing thinking depth. Less thinking leads to higher accuracy for WebSailor.

7.3 Ablation Study

To investigate how different knowledge and rules in HSCODECOMP affect performance, we perform ablation studies to isolate their contribution.

Inefficacy of Expert-written Decision Rules. Intuitively, providing expert-written hierarchical decision rules should guide the agent's reasoning.

Agent System	Type	10-digit Accuracy		
		w/o DR	w/ DR	Δ
SmolAgents	VLM	**46.83**	43.83	-3.00
WebSailor	LLM	**35.44**	35.34	-0.10

Table 5. The ablation study on expert-written **D**ecision **R**ules (**DR**). All systems use GPT-5 as the backbone.

However, Table 5 reveals that injecting these explicit guidelines degrades performance. For example, SmolAgent with GPT-5 degrades from 46.83% → 43.83%. We attribute this failure to an **Rule-Grounding Gap**: unlike human experts who flexibly adapt high-level principles to specific contexts, current agents struggle to ground noisy, concrete product profile into abstract rule boundaries. Consequently, to prevent such reasoning degradation, we exclude these explicit decision rules from the default evaluation setup.

Hierarchical Tariff Rules are Essential. Removing the Tariff Rules results in a significant performance drop. SmolAgent with GPT-5 degrades from 46.83% → 36.73%. It confirms that HSCode prediction is fundamentally a rule-centric task rather than a retrieval-centric one. Agents cannot rely solely on open-web knowledge or model's internal knowledge; they must strictly adhere to the specific, hierarchical tariff rules to resolve fine-grained classification paths.

Supplementary Role of Knowledge Base. Access to the CROSS knowledge base yields only a marginal performance benefit (SmolAgent w/ GPT-5, 46.53% → 46.83%), indicating that while historical precedents provide useful auxiliary context for resolving specific ambiguities, they remain supplementary to the core reasoning process.

Visual Signal are Helpful. Table 1 and Table 6 show that incorporating product images yields a significant gain. For example, SmolAgent with GPT-5 improves from 42.72% → 46.83% when visual capability is activated. Case studies in Appendix L show that visual attributes that are not present in the textual description (such as material and surface features) are critical for classification.

7.4 Qualitative Analysis: Failure Taxonomy

To systematically diagnose agent failures, we classify the identified error patterns into three categorizations: Perception, Reasoning, and Reflection.

Backbone Model	10-digit Accuracy		
	w/o Image	w/ Image	Δ
GPT-5	42.72	**46.83**	**+4.11**
Claude 4 Sonnet	33.70	**34.65**	**+0.95**
GPT-4o	22.03	**22.31**	**+0.28**

Table 6. Ablation study of visual signal.

Perception & Grounding Failures. Two kinds of errors often originate before reasoning: (1) **Information Misprocessing**: Agents overlook or misinterpret product details. For instance, in Figure 15, the agent becomes fixated on marketing buzzwords ("INS style", "Aesthetic") while overlooking the decisive technical specification ("Material: PVC"), leading to a misclassification based on style rather than substance; (2) **Lack of Domain Knowledge:** As seen in Figure 16, models may misidentify *silicone* products as *rubber* (Chapter 40) instead of *plastics* (Chapter 39), lacking the correct domain-specific ontology to distinguish material properties essential for customs classification.

Reasoning & Planning Failures. Even when input is correctly perceived, agents struggle to navigate the logical decision on hierarchical rules: (1) **Premature Decisions:** This is a planning failure where agents lock onto a high-level category too early, as seen in Table 14; (2) **Wrong Rule Application:** As detailed in Figure 17, a prevalent failure is the agent's inability to dynamically determine the correct rule precedence given the product description; and (3) **Reasoning Hallucination:** Table 13 reveal that agents generate plausible but factually incorrect reasoning.

Reflection. Finally, agents frequently conduct wrong self-correction: agents predict correct HSCodes initially but revise them incorrectly through excessive critique (Table 13).

7.5 Per-category Performance Analysis

We analyze two critical distributions across the 32 first-level product categories to diagnose whether performance correlates with data distribution: (1) **Challenging Product Distribution (CID)**: the distribution of hard instances where all baselines failed; (2) **Average Performance Distribution (APD)**: the distribution of average 10-digit accuracy across all baselines. Figure 5 reveals two key insights: (1) The CID indicates that the most challenging products are concentrated in long-tail categories, such as *Novelty & Special Use* (1.3%) and *Men's Clothing* (2.2%). These categories, characterized by non-standard descriptions and sparse examples, expose the agents' inability to generalize rules to data-sparse domains; (2) The APD shows that average accuracy across most product categories remains below 25%, with only *Hair Extensions & Wigs* achieving a relatively high accuracy of 47%. Importantly, even dominant categories like *Jewelry & Accessories* (13.1%), *Home & Garden* (16.8%) and *Tools* (6.2%) still stays below 18%. These findings highlight the challenging nature of HSCODECOMP.

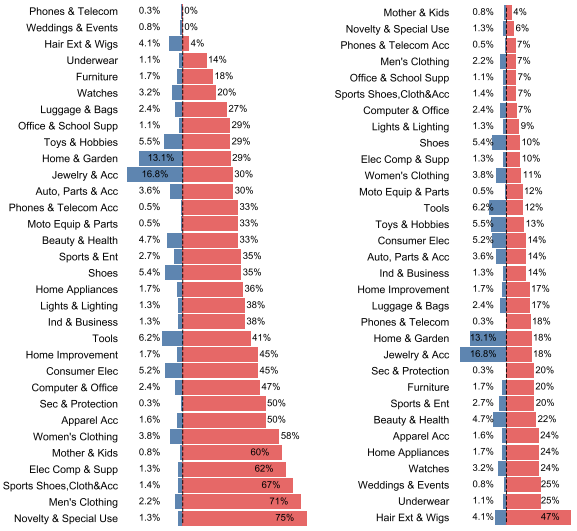


Figure 5. Category distribution on the left blue bars. **Left:** Challenging Product Distribution (CID). **Right:** Average Performance Distribution (APD).

8 Conclusion

We identified and addressed the critical gap in evaluating agents in hierarchical rule applications. To address this gap, we introduced HSCODECOMP, the first realistic and expert-level benchmark designed to assess agents for multi-hop reasoning with hierarchical tariff rules in e-commerce domain. Our extensive evaluation revealed a substantial performance gap between current state-of-the-art agents (46.8%) and human experts (95.0%), highlighting that hierarchical rule application remains a significant challenge for existing agent architectures. We will release the HSCODECOMP to accelerate research in this crucial capability for real-world agent deployment.

Limitation

Temporal Sensitivity. International trade policies are dynamic. As noted in Section 4 and Appendix I, our ground truth is anchored to a specific timestamp (June 30, 2025), which is a standard implementation in the research community (Wong et al., 2025). While this limits our ability to evaluate an agent's real-time temporal adaptability, it does not compromise the validity of the benchmark. By strictly anchoring the evaluation to a specific version of the tariff schedule[2], we effectively isolate the agent's reasoning capability from the volatility of external knowledge, ensuring the benchmark measures the agent's ability to interpret and apply a given set of rules, regardless of their temporal version.

Trade-off Between Benchmark Quality and Scale. Diverging from the prevalence of crowdsourced or synthetic benchmarks (Zhou et al., 2025; Phan et al., 2025; Wu et al., 2025), HSCODECOMP prioritizes label quality over scale. The average hourly wage of annotators is $\approx$ \$34.6/hour. The dataset size (632 entries) reflects a deliberate trade-off necessitated by the resource-intensive nature of verification by domain experts. While our statistical analysis (Appendix E) confirms that this sample size provides sufficient power for stable model differentiation, we acknowledge that the benchmark may not exhaustively capture the extreme long-tail of global trade, particularly those with zero-shot distributions.

The Scope of Tariff Rules. Currently, HSCODECOMP is exclusively scoped to the United States Harmonized Tariff Schedule (HTSUS). While the HS system is globally standardized at the 6-digit level, classifications at the 8-to-10 digit level remain country-specific. Consequently, our benchmark evaluates agents primarily on US-centric legal frameworks and English-language descriptions. Future work will need to expand to multilingual and multi-jurisdictional scenarios (e.g., EU or China customs) to assess cross-border regulatory generalization.

References

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, and Oren Tropp et al. 2025. Llama-nemotron: Efficient reasoning models. *Preprint*, arXiv:2505.00949.

Kaiyuan Chen, Yixin Ren, Yang Liu, Xiaobo Hu, Haotong Tian, Tianbao Xie, Fangfu Liu, Haoye Zhang, Hongzhang Liu, Yuan Gong, Chen Sun, Han Hou, Hui Yang, James Pan, Jianan Lou, Jiayi Mao, Jizheng Liu, Jinpeng Li, Kangyi Liu, and 14 others. 2025a. xbench: Tracking agents productivity scaling with profession-aligned real-world evaluations. *Preprint*, arXiv:2506.13651.

Shan Chen, Pedro Moreira, Yuxin Xiao, Sam Schmidgall, Jeremy Warner, Hugo Aerts, Thomas Hartvigsen, Jack Gallifant, and Danielle S. Bitterman. 2025b. Medbrowsecomp: Benchmarking medical deep research and computer use. *Preprint*, arXiv:2505.14963.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, and Zhibin Gou et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Tianqing Fang, Zhisong Zhang, Xiaoyang Wang, Rui Wang, Can Qin, Yuxuan Wan, Jun-Yu Ma, Ce Zhang, Jiaqi Chen, Xiyun Li, Hongming Zhang, Haitao Mi, and Dong Yu. 2025. Cognitive kernel-pro: A framework for deep research agents and agent foundation models training. *Preprint*, arXiv:2508.00414.

Sina Gholamian, Gianfranco Romani, Bartosz Rudnikowicz, and Laura Skylaki. 2024. Llm-based robust product classification in commerce and compliance. *arXiv preprint arXiv:2408.05874*.

Andrew Grainger. 2024. Customs tariff classification and the use of assistive technologies. *World Customs Journal*, 18(1):3–32.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Re, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, and et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kotaro Hara, Abi Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey Bigham. 2017. A data-driven analysis of workers' earnings on amazon mechanical turk. *Preprint*, arXiv:1712.05796.

Liang Hu, Jianpeng Jiao, Jiashuo Liu, Yanle Ren, Zhoufutu Wen, Kaiyuan Zhang, Xuanliang Zhang, Xiang Gao, Tianci He, Fei Hu, Yali Liao, Zaiyuan Wang, Chenghao Yang, Qianyu Yang, Mingren Yin, Zhiyuan Zeng, Ge Zhang, Xinyi Zhang, Xiying Zhao, and 4 others. 2025a. Finsearchcomp: Towards a realistic, expert-level evaluation of financial search and reasoning. *Preprint*, arXiv:2509.13160.

[2] https://www.ewtp.com/web

Mengkang Hu, Yuhang Zhou, Wendong Fan, Yuzhou Nie, Bowei Xia, Tao Sun, Ziyu Ye, Zhaoxuan Jin, Yingru Li, Qiguang Chen, Zeyu Zhang, Yifeng Wang, Qianshuo Ye, Bernard Ghanem, Ping Luo, and Guohao Li. 2025b. Owl: Optimized workforce learning for general multi-agent assistance in real-world task automation. *Preprint*, arXiv:2505.23885.

Kung-Hsiang Huang, Akshara Prabhakar, Sidharth Dhawan, Yixin Mao, Huan Wang, Silvio Savarese, Caiming Xiong, Philippe Laban, and Chien-Sheng Wu. 2025. Crmarena: Understanding the capacity of llm agents to perform professional crm tasks in realistic environments. In *Proceedings of NAACL 2025 (Long Papers)*, Albuquerque, New Mexico. Association for Computational Linguistics.

Mandar Joshi and 1 others. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*.

Bryce Judy. 2024. Benchmarking harmonized tariff schedule classification models. *arXiv preprint arXiv:2412.14179*.

Eunji Lee, Sihyeon Kim, Sundong Kim, Soyeon Jung, Heeja Kim, and Meeyoung Cha. 2024. Explainable product classification for customs. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–24.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, Wuyue Wang, Yiqun Liu, and Minlie Huang. 2025a. Legalagentbench: Evaluating llm agents in legal domain. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria. Association for Computational Linguistics.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong Jiang, Ming Yan, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025b. Websailor: Navigating super-human reasoning for web agent. *Preprint*, arXiv:2507.02592.

Mengjie Liao, Lei Huang, Jian Zhang, Luona Song, and Bo Li. 2024. Enhanced hs code classification for import and export goods via multiscale attention and ernie-bilstm. *Applied Sciences*, 14(22):10267.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025. Inference-time scaling for generalist reward modeling. *Preprint*, arXiv:2504.02495.

Yougang Lyu, Xiaoyu Zhang, Lingyong Yan, Maarten de Rijke, Zhaochun Ren, and Xiuying Chen. 2025. Deepshop: A benchmark for deep research shopping agents. *Preprint*, arXiv:2506.02839.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023a. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023b. Gaia: a benchmark for general ai assistants. *Preprint*, arXiv:2311.12983.

Ralph Peeters, Aaron Steiner, Luca Schwarz, Julian Yuya Caspary, and Christian Bizer. 2025. Webmall – a multi-shop benchmark for evaluating web agents. *arXiv preprint arXiv:2508.13024*.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, and 1 others. 2025. Humanity's last exam. Center for AI Safety and Scale AI (whitepaper / dataset).

Aymeric Roucher, Albert Villanova del Moral, Thomas Wolf, Leandro von Werra, and Erik Kaunismäki. 2025. 'smolagents': a smol library to build great agentic systems. `https://github.com/huggingface/smolagents`.

Albert Sadowski and Jarosław A. Chudziak. 2025. Explainable rule application via structured prompting: A neural-symbolic approach. *Preprint*, arXiv:2506.16335.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In *ICLR*.

Shubham, Avinash Arya, Subarna Roy, and Sridhar Jonnala. 2022. An ensemble-based approach for assigning text to correct harmonized system code. *arXiv preprint arXiv:2211.04313*.

Sédrick Stassin, Otmane Amel, Sidi Mahmoudi, and Xavier Siebert. 2023. Similarity versus supervision: Best approaches for hs code prediction.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, and Ruijue Chen et al. 2025. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

George Thomas, Alex J. Chan, Jikun Kang, Wenqi Wu, Filippos Christianos, Fraser Greenlee, Andy Toulis, and Marvin Purtorab. 2025. Webgames: Challenging general-purpose web-browsing ai agents. In *arXiv preprint arXiv:2502.18356*.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *Preprint*, arXiv:2504.12516.

Ryan Wong, Jiawei Wang, Junjie Zhao, Li Chen, Yan Gao, Long Zhang, Xuan Zhou, Zuo Wang, Kai Xiang, Ge Zhang, Wenhao Huang, Yang Wang, and Ke Wang. 2025. Widesearch: Benchmarking agentic broad info-seeking. *Preprint*, arXiv:2508.07999.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. 2025. Webwalker: Benchmarking llms in web traversal. *Preprint*, arXiv:2501.07572.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, and Bowen Yu et al. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Hong Yao, Zipeng Jing, Tianshi Liu, Eric J Wong, Chunyuan Hong, Is Wang, Wen-Loong Wang, Yalda Talebirad, Rui Wang, Zhaowei Chen, and 1 others. 2024. Webvoyager: Building an end-to-end web agent with large multimodal models. In *ICLR*.

Shunyu Yao, Howard Chen, John Yang, and Karthik R. Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. In *NeurIPS*.

Ailing Yu, Lan Yao, Jingnan Liu, Zhe Chen, Jiajun Yin, Yuan Wang, Xinhao Liao, Zhiling Ye, Ji Li, Yun Yue, Hansong Xiao, Hualei Zhou, Chunxiao Guo, Peng Wei, Junwei Liu, and Jinjie Gu. 2025a. Medresearcher-r1: Expert-level medical deep researcher via a knowledge-informed trajectory synthesis framework. *Preprint*, arXiv:2508.14880.

Ailing Yu, Lan Yao, Jingnan Liu, Zhe Chen, Jiajun Yin, Yuan Wang, Xinhao Liao, Zhiling Ye, Ji Li, Yun Yue, Hansong Xiao, Hualei Zhou, Chunxiao Guo, Peng Wei, Junwei Liu, and Jinjie Gu. 2025b. Medresearcher-r1: Expert-level medical deep researcher via a knowledge-informed trajectory synthesis framework. *arXiv preprint arXiv:2508.14880*.

Chengyue Yu, Siyuan Lu, Chenyi Zhuang, Dong Wang, Qintong Wu, Zongyue Li, Runsheng Gan, Chunfeng Wang, Siqi Hou, Gaochi Huang, Wenlong Yan, Lifeng Hong, Aohui Xue, Yanfeng Wang, Jinjie Gu, David Tsai, and Tao Lin. 2025c. Aworld: Orchestrating the training recipe for agentic ai. *Preprint*, arXiv:2508.20404.

Wentao Zhang, Liang Zeng, Yuzhen Xiao, Yongcong Li, Ce Cui, Yilei Zhao, Rui Hu, Yang Liu, Yahui Zhou, and Bo An. 2025. Agentorchestra: A hierarchical multi-agent framework for general-purpose task solving. *Preprint*, arXiv:2506.12508.

Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, Yuxin Gu, Sixin Hong, Jing Ren, Jian Chen, Chao Liu, and Yining Hua. 2025. Browsecomp-zh: Benchmarking web browsing ability of large language models in chinese. *Preprint*, arXiv:2504.19314.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2023. Webarena: A realistic web environment for building autonomous agents. In *arXiv preprint arXiv:2307.13854*.

King Zhu, Hanhao Li, Siwei Wu, Tianshun Xing, Dehua Ma, Xiangru Tang, Minghao Liu, Jian Yang, Jiaheng Liu, Yuchen Eleanor Jiang, Changwang Zhang, Chenghua Lin, Jun Wang, Ge Zhang, and Wangchunshu Zhou. 2025. Scaling test-time compute for llm agents. *Preprint*, arXiv:2506.12928.

A Ethics Statement

This research adheres to strict ethical guidelines regarding data privacy and fair labor. The dataset is fully anonymized and contains no personally identifiable information. The hourly wage of our human annotators is over 34.6 USD, which is much higher than average hourly wage 3.13 USD on Amazon Mechanical Turk (Hara et al., 2017). This remuneration structure was designed to provide a fair and competitive wage, acknowledging the expertise and effort required for this task and ensuring that contributors were rewarded appropriately for their work.

B Reproducibility statement

We are committed to the principles of reproducible research. Accordingly, all necessary materials, including code, benchmark dataset and other related resources will be publicly released to promote the development of the deep search agents.

C The Use of Large Language Models

In preparing this manuscript, Qwen-Max and ChatGPT were used solely as a writing assistant to improve grammar and clarity. The LLMs was not used for generating code, concepts, or any part of the core research methodology.

D Dataset distribution

In this section, we provide following detailed data statistics of our proposed HSCODECOMP: (1) Total unique 10-digit codes is 352; (2) Sample distribution per HSCode is shown in Table 7. Since 72.73% of HSCodes appear only once in our dataset, HSCODECOMP is not intentionally enriched for tricky or dispute-prone items; and (3) HSCode subcategory Hierarchical Coverage is shown in Table 8. Besides, Figure 6 presents the distributions of first-level product categories (left) and HSCode chapter categories (right), which closely mirror real-world product distributions. This alignment confirms that HSCODECOMP accurately reflects practical international trade scenarios, ensuring that model performance evaluations reliably generalize to real-world applications. In summary, our sampling strategy for constructing HSCODECOMP directly follows sales volume and category distributions from a major global E-Commerce platform. This approach ensures the benchmark reflects actual production distributions.

Samples per Code	Count	Percentage
1	256	72.73%
2	58	16.48%
3	16	4.55%
4	4	1.14%
5	5	1.42%
6+	13	3.69%

Table 7. Sample Distribution per HSCode

HSCode Subcategory	Count
2-digit (chapters)	27
4-digit (headings)	151
6-digit (subheadings)	268
8-digit (country-specific)	322
10-digit (country-specific)	352

Table 8. HSCode Hierarchical Coverage

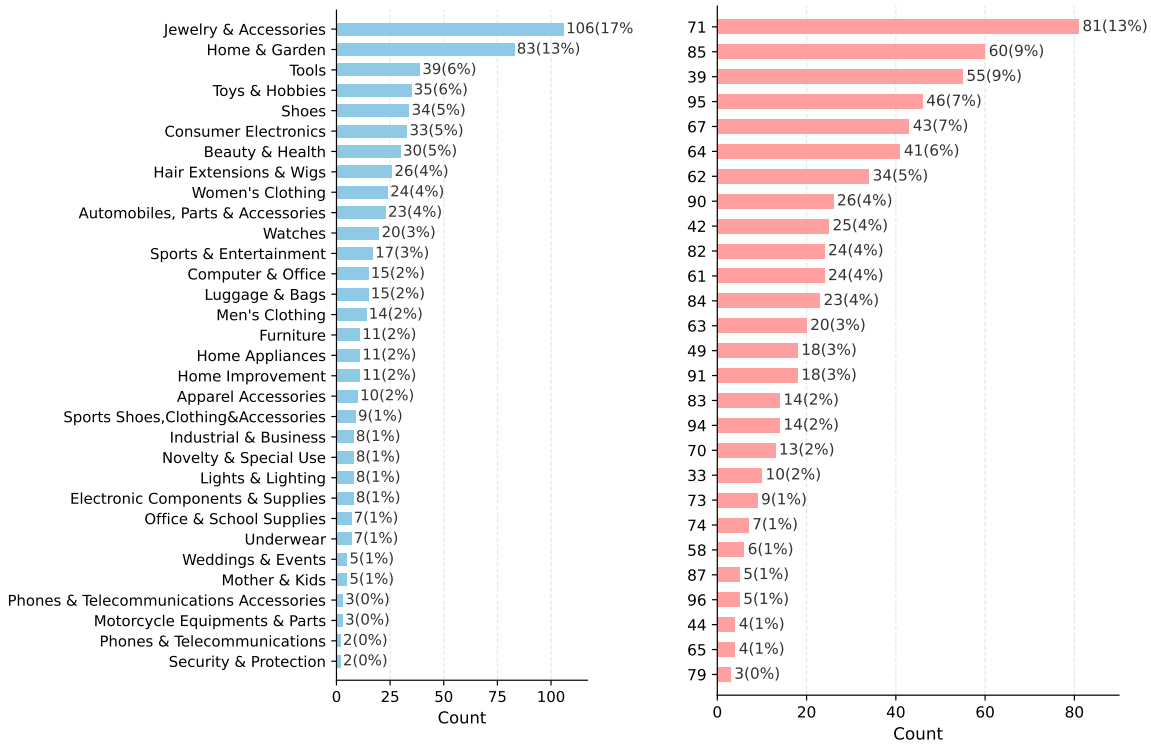
E Statistical Stability Analysis

A primary concern in constructing expert-level benchmarks is balancing annotation cost with statistical significance. To validate the robustness of HSCODECOMP (632 samples), we conducted a bootstrap analysis of the 95% Confidence Intervals (CI) across varying sample sizes for seven representative agent systems.

Confidence Interval Convergence. We calculated the 95% CI widths for sample sizes ranging from 200 to 632. As detailed in Table 9, the evaluation exhibits clear diminishing returns in CI width reduction beyond 600 samples.

For the best-performing system (`SmolAgents` with VLM), the CI width decreases by only **0.21%** when expanding the dataset from 600 to 632 samples (reducing from 7.99% to 7.78%). This marginal improvement indicates that the evaluation has reached statistical equilibrium at the current dataset size. Further expansion would yield negligible gains in statistical precision.

Comparison with Community Standards. Despite the smaller absolute number of samples compared to automated benchmarks, our statistical reliability aligns with or exceeds established agentic benchmarks. For instance, top models on the GAIA benchmark exhibit 95% confidence intervals of approximately $\pm 8.2\%$ (OpenAI GPT-4 Turbo) to $\pm 8.7\%$ (Claude 3 Sonnet). In comparison, HSCODECOMP yields narrower or comparable confidence intervals (e.g., $\pm 7.78\%$ width for our best agent), demonstrating that 632 expert-validated instances are sufficient to differentiate



(a) First-level product categories

(b) HSCodes distribution on chapter

Figure 6. Distributions of the first-level product category and HSCode chapter categories.

model performance reliably.

Resource Allocation Efficiency. It is also important to note the high cost of expert annotation for hierarchical rule application. As detailed in our Ethics Statement, annotators are domain professionals paid significantly above market rates ($34.6/hour). Given the statistical stability demonstrated in Table 9, expanding the dataset further provides minimal scientific value relative to the substantial resource cost required for expert validation.

F Semantic Distribution of Tariff Rules

To assess whether the HSCode taxonomy exhibits clear semantic separation, we generate embeddings of the official English titles and notes for all HS chapters and sections using a sentence embedding model[3]. We then apply t-SNE to project these embeddings into two dimensions for visualization. As shown in Figure 7, each point represents a chapter, while each star marks a section's centroid. The visualization reveals significant semantic overlap between adjacent sections: numerous chapters appear closer to neighboring section centroids than to their own section's centroid, and section centroids themselves form overlapping clusters rather than distinct groupings. This pattern indicates that the semantic structure of hierarchical tariff rules lacks clear boundaries—adjacent sections frequently share similar vocabulary and concepts (*e.g.*, distinctions between raw materials and finished goods, or between component parts and complete articles).

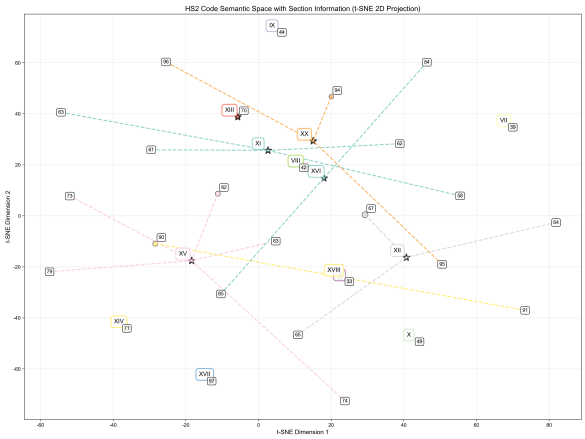


Figure 7. The semantic map of HS chapter titles and notes.

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Model	200	300	400	500	600	632 (Full)	$\Delta_{600\to632}$
SmolAgents (VLM)	[42.6, 47.6] (11.17%)	[41.6, 51.4] (9.80%)	[46.1, 55.9] (8.75%)	[36.4, 55.4] (13.85%)	[43.3, 51.3] (7.99%)	[42.9, 50.7] (7.78%)	0.21%
AWorld	[37.1, 50.9] (13.76%)	[35.1, 43.7] (10.91%)	[37.9, 42.1] (9.70%)	[31.2, 47.6] (8.57%)	[36.2, 44.1] (7.85%)	[37.5, 45.1] (7.68%)	0.17%
WebSailor	[29.3, 47.8] (13.63%)	[31.1, 40.4] (10.61%)	[34.2, 37.5] (9.39%)	[27.4, 38.0] (8.27%)	[29.6, 37.1] (7.54%)	[31.7, 39.2] (7.46%)	0.08%
GPT-5 (VLM)	[23.6, 35.8] (12.70%)	[26.7, 32.1] (10.23%)	[23.5, 36.4] (9.08%)	[24.3, 33.8] (7.89%)	[25.2, 32.5] (7.25%)	[25.7, 32.8] (7.09%)	0.16%
GPT-5 (LLM)	[24.1, 36.1] (12.58%)	[22.7, 31.9] (9.10%)	[22.9, 33.1] (10.16%)	[26.9, 35.3] (7.87%)	[25.0, 32.3] (7.24%)	[25.4, 32.5] (7.07%)	0.17%
Claude Sonnet 4 (VLM)	[7.5, 14.5] (7.08%)	[7.7, 16.5] (9.01%)	[8.6, 14.9] (6.31%)	[7.5, 13.1] (5.35%)	[8.5, 13.5] (5.01%)	[9.3, 14.4] (5.04%)	-0.03%

Note: Values in brackets represent the 95% confidence interval range. Values in parentheses represent the CI width.

Table 9. Analysis of 95% Confidence Interval (CI) widths across varying sample sizes. The "Δ" column represents the reduction in CI width when increasing samples from 600 to 632. The results show that the CI width stabilizes around 600 samples.

G Supplementary Experimental Results

G.1 Conditional Performance

As shown in Table 10, we provide the conditional accuracy of partial models on Chapter, Heading, Sub-heading and Country-specific sub-categories code classification. The experimental results reveal three key insights: (1) The 2-4 digit heading level shows highest average accuracy (75.16), since its tariff rules are most clearly defined; (2) Intermediate performance drops at the 4-6 digit subheading level due to increased the number of rules and their vague descriptions in tariff rules; and (3) The steepest decline occurs at the country-specific 6-10 digit levels, where rules become highly contextual and country-specific. This hierarchical performance pattern directly supports our main finding: current agents struggle with the complex, multi-step reasoning required for precise rule application at deeper hierarchical levels. The significant accuracy drop from 64.8% (6-digit) to 35.3% (10-digit) demonstrates that correctly navigating the initial hierarchical steps does not guarantee success in the final classification, highlighting the cumulative reasoning challenges in hierarchical rule application.

G.2 Detailed Ablation Study on Rules

We provide the complete ablation study on used domain-specific knowledge and rules. As shown in Table 11, we remove each used rules or knowledge in WebSailor and Smolagent open-source agent frameworks for solving tasks in HSCODECOMP: (1) **DR**: Human-written Decision Rules; (2) **Tariff**: hierarchical tariff rules; and (3) **CROSS** ruling database. Experimental results reveal three critical insights: (1) Human-written decision rules impair performance: As shown in Table 3 of our paper, adding decision rules (DR) decreases accuracy by 3.00% for Smolagent (GPT-5), demonstrating current agent systems cannot effectively utilize hierarchical rule structures; (2) Tariff rules are essential and contributes most: Removing tariff rules causes a severe 10.10% drop in accuracy. Without tariff rules, agents have to solve the task using its internal knowledge, leading to outdated and hallucination problems; and (3) CROSS provides moderate benefit: The 4.61% accuracy drop when removing CROSS indicates it offers supplementary but non-essential information. These results confirm that tariff rules are the most critical knowledge component, while human-written decision rules present the most significant challenge for current agent architectures to utilize effectively.

G.3 Experiments on Backbone Models

To investigate how the backbone LLMs affect the performance of the agent system, we conduct more detailed ablation study on four open-source agent systems, replacing the original GPT-5 backbone with Gemini 2.5 Pro. The experimental results in Table 12 indicate that GPT-5 achieves consistently better performance than the advanced Gemini 2.5 pro model on these open-source agent systems.

G.4 Ablation Study on Webpage Visit Tool

Augmenting agents with full webpage navigation, for example, visiting real-world e-commerce pages with marketing noise, decrease performance (Smo-

Baselines	Model Type	HSCode Prediction Accuracy			
		2-digit Chapter Acc.	4-digit Heading Acc.	6-digit Sub-heading Acc.	10-digit Country-specific Acc.
LLM/VLM-Only					
GPT-5	VLM	82.12	86.32	84.60	48.81
Gemini-2.5-PRO	VLM	82.28	86.34	83.08	41.02
Claude Sonnet 4	VLM	78.80	81.32	70.61	24.82
Kimi-K2	LLM	78.01	79.52	71.18	27.59
DeepSeek-V3	LLM	77.06	70.63	59.31	20.11
O3-mini	LLM	77.22	72.74	43.67	5.18
Open-source Agent System (GPT-5 Backbone)					
SmolAgents	VLM	82.06	87.81	86.57	75.07
Aworld	LLM	82.28	85.57	84.05	69.79
Agentorchestra	LLM	82.12	86.13	85.45	68.33
OWL	LLM	72.63	85.19	83.37	72.39
WebSailor	LLM	81.64	86.43	81.16	61.88
Cognitive Kernel	LLM	80.06	86.37	78.94	48.40
Average	-	73.80	75.16	64.78	35.32

Table 10. The conditional performance on Chapter, Heading, Sub-heading and Country-specific codes classifiction in our proposed HSCODECOMP.

Configuration	10-digit Acc.
Full Framework	43.83%
(w/ Tariff, Rules, CROSS)	
w/o Decision Rules (DR)	**46.83%**
w/o Tariff Rules	36.73%
w/o CROSS Knowledge	46.53%

Table 11. **Ablation study on Knowledge Sources.** We evaluate the impact of removing distinct knowledge components from the SmolAgent (GPT-5) backbone.

lAgent GPT-5 LLM: 42.72% → 42.09%). These webpage content often overwhelms the key information, misleading the agents. Therefore, we remove the webpage visit tool for all open-source agents as the default setup.

H Improvement Gain from Agents

This waterfall in Figure 8 chart reveals that the superior performance of SmolAgent (GPT-5) are primarily from reducing the outdated and hallucination failures, with 56 corrected samples in HSCODBENCH. Besides, as shown in Figure 9, it can be found that the rate of outdated and hallucination are significantly reduced in SmolAgent baseline.

I Implementing Details

Evaluation Details. All baseline methods are equipped with search tools to access the CROSS database, hierarchical tariff rules, and other related resources, including human-written knowledge bases and hierarchical decision rules. The

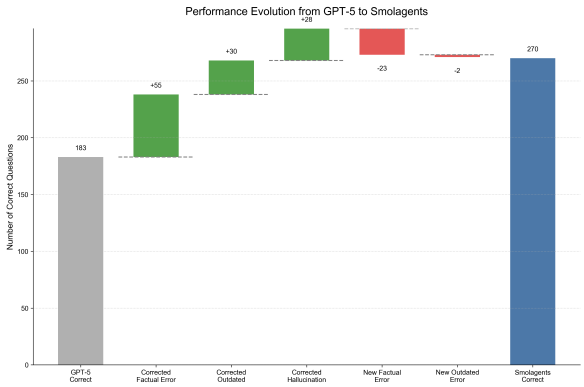


Figure 8. Details of performance gain and loss comparing GPT-5 and Smolagents.

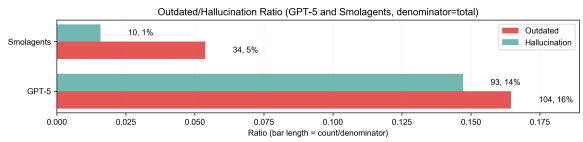


Figure 9. Outdated or hallucination ratio happended on GPT-5 and Smolagent. The numbers are number and ratio of the phenomenon.

Backbone LLM	HSCode Prediction Accuracy				
	2-digit	4-digit	6-digit	8-digit	10-digit
SmolAgent Framework					
GPT-5	**82.28**	**70.89**	**59.81**	**49.05**	**42.72**
Gemini-2.5-Pro	82.19	69.48	57.87	44.04	34.49
Claude 4 Sonnet	80.69	67.09	54.11	42.25	33.70
Qwen-Max	77.34	63.23	42.47	26.62	17.43
Aworld Framework					
GPT-5	**82.28**	**70.41**	**59.18**	**48.58**	**41.30**
Gemini 2.5 Pro	79.55	66.97	54.70	38.79	29.24
WebSailor Framework					
GPT-5	**81.64**	**70.56**	**57.27**	**43.98**	**35.44**
Gemini 2.5 Pro	78.79	67.58	56.21	42.27	31.21
AgentOrchestra Framework					
GPT-5	82.12	**70.73**	**60.44**	**47.78**	**41.30**
Gemini 2.5 Pro	**82.27**	69.39	56.97	41.36	30.61

Table 12. **Ablation study of backbone models across different open-source agent systems.** Results indicate that GPT-5 consistently outperforms other models across all frameworks.

temperature and context window size of LLMs and agents are set to their default configurations. Moreover, as described in Section 6, the hierarchical decision rules, and webpage visit tool are not used during evaluation, since they do not improve the performance of open-source and closed-source agents. The multi-modal product images are used for open-source agents for SmolAgents. The hierarchical decision rules used in our prompts are illustrated in Figure I.1. The hierarchical tariff rules in the eWTP is shown in Figure 10. It can be found that the red boxes highlight implicit logical relationships in the tariff rules, such as *excluding articles of heading 8593* and *with the machines of heading 8501 or 8502*. The blue boxes highlight vague descriptions in the tariff rules, such as ... *for example* ... and ... *such as* .... These cases demonstrate that rule boundaries are ambiguous, posing significant challenges for accurate rule application by the agent. Moreover, the U.S. Customs Rulings Online Search System (CROSS) interface is shown in Figure 11. As illustrated, the CROSS website contains not only correct precedent results for product HS Codes but also numerous revoked precedents, requiring the agent to carefully evaluate information reliability. Additionally, since the precedent information is presented as plain text emails, the agent must effectively utilize contextual information to perform accurate reasoning.

Knowledge Alignment. All these knowledge resources in HSCODECOMP: hierarchical rules come directly from official US HTS legal notes included in eWTP, and the CROSS database contains authentic US Customs rulings. Human-written decision rules are also developed by human experts with US compliance experience.

Timestamp-based Evaluation. To ensure the reproducibility of our proposed HSCODECOMP dataset, all our dataset, knowledge and experimental setup are tied to the specific timestamp—June 30, 2024. This is a wide-used community standards of established benchmarks like BrowseComp (Wei et al., 2025) and WideSearch (Wong et al., 2025).

I.1 Infeasibility of Flat HSCode Prediction

A potential alternative to our hierarchical agentic framework is a "flat" classification approach, where a model directly predicts the 10-digit HSCode from the product description. To evaluate the feasibility of this approach, we conducted a preliminary study using a `Qwen3-8B` model fine-tuned on a dataset of 175,000 historical customs rulings.

Despite the substantial training data, the flat prediction model achieved a 10-digit accuracy of only **0.16%**. This catastrophic failure is attributed to the extreme long-tail distribution of the 10-digit classification space (which contains over 19,000 valid codes in the Harmonized Tariff Schedule of the United States (HTSUS)). The model struggled

Section XVI Machinery and mechanical appliances; electrical equipment; parts thereof; sound recorders and reproducers, television image and sound recorders and reproducers, and parts and accessories of such articles Notes

84 Nuclear reactors, boilers, machinery and mechanical appliances; parts thereof

85 Electrical machinery and equipment and parts thereof; sound recorders and reproducers, television image and sound recorders and reproducers, and parts and accessories of such articles

8501 Electric motors and generators (excluding generating sets):

8502 Electric generating sets and rotary converters:

850300 Parts suitable for use solely or principally with the machines of heading 8501 or 8502:

8504 Electrical transformers, static converters (for example, rectifiers) and inductors; parts thereof:

8505 Electromagnets; permanent magnets and articles intended to become permanent magnets after magnetization; electromagnetic or permanent magnet chucks, clamps and similar holding devices; electromagnetic couplings, clutches and brakes; electromagnetic lifting heads; parts thereof:

8506 Primary cells and primary batteries; parts thereof:

8507 Electric storage batteries, including separators therefor, whether or not rectangular (including square); parts thereof:

8508 Vacuum cleaners; parts thereof:

8509 Electromechanical domestic appliances, with selfcontained electric motor, other than vacuum cleaners of heading 8508; parts thereof:

8510 Shavers, hair clippers and hairremoving appliances, with selfcontained electric motor; parts thereof:

8511 Electrical ignition or starting equipment of a kind used for sparkignition or compressionignition internal combustion engines (for example, ignition magnetos, magnetodynamos, ignition coils, spark plugs and glow plugs, starter motors); generators (for example, dynamos, alternators) and cutouts of a kind used in conjunction with such engines; parts thereof:

8512 Electrical lighting or signaling equipment (excluding articles of heading 8539), windshield wipers, defrosters and demisters, of a kind used for cycles or motor vehicles; parts thereof:

8513 Portable electric lamps designed to function by their own source of energy (for example, dry batteries, storage batteries, magnetos), other than lighting equipment of heading 8512; parts thereof:

8514 Industrial or laboratory electric furnaces and ovens (including those functioning by induction or dielectric loss); other industrial or laboratory equipment for the heat treatment of materials by induction or dielectric loss; parts thereof:

8515 Electric (including electrically heated gas), laser or other light or photon beam, ultrasonic, electron beam, magnetic pulse or plasma arc soldering, brazing or welding machines and apparatus, whether or not capable of cutting; electric machines and apparatus for hot spraying of metals or cermets; parts thereof:

8516 Electric instantaneous or storage water heaters and immersion heaters; electric space heating apparatus and soil heating apparatus; electrothermic hairdressing apparatus (for example, hair dryers, hair curlers, curling tong heaters) and hand dryers; electric flatirons; other electrothermic appliances of a kind used for domestic purposes; electric heating resistors, other than those of heading 8545; parts thereof:

8517 Telephone sets, including smartphones and other telephones for cellular networks or for other wireless networks; other apparatus for the transmission or reception of voice, images or other data, including apparatus for communication in a wired or wireless network (such as a local or wide area network), other than transmission or reception apparatus of heading 8443, 8525, 8527 or 8528; parts thereof:

8518 Microphones and stands therefor; loudspeakers, whether or not mounted in their enclosures; headphones and earphones, whether or not combined with a microphone, and sets consisting of a microphone and one or more loudspeakers; audiofrequency electric amplifiers; electric sound amplifier sets; parts thereof:

8519 Sound recording or reproducing apparatus:

8521 Video recording or reproducing apparatus, whether or not incorporating a video tuner:

Figure 10. The case of the hierarchical tariff rules. The red boxes highlight implicit logical relationships, such as exclusion clauses ("excluding...") and cross-references ("with the machines of heading 8501 or 8502"). The blue boxes indicate vague decision boundaries characterized by open-ended illustrative examples (e.g., "for example," "such as"), requiring agents to perform semantic interpretation beyond literal matching.

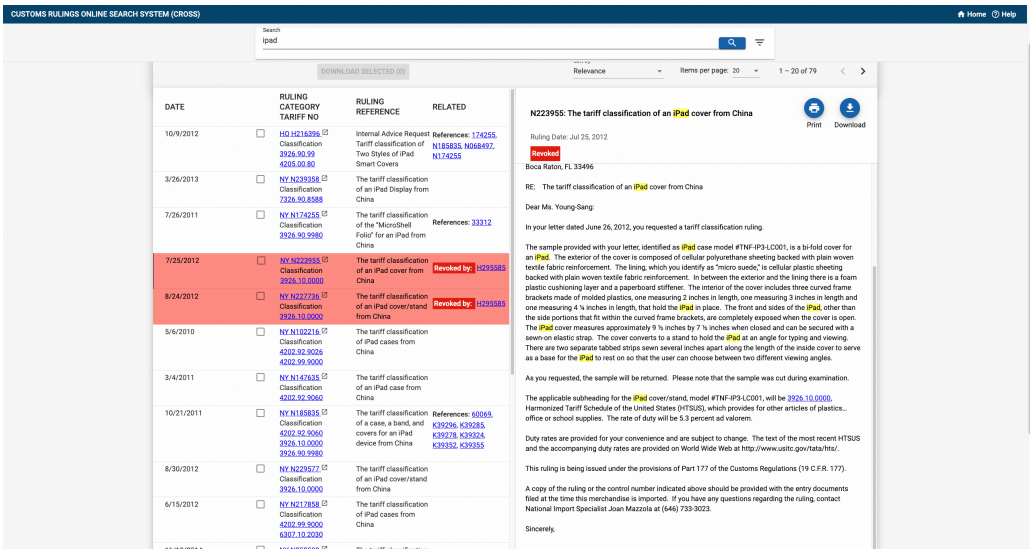


Figure 11. The case of CROSS knowledge database that contains the products rulings.

to memorize the specific mappings for rare codes and lacked the reasoning capability to navigate the legal boundaries between similar codes.

This result confirms that a flat classification baseline is practically infeasible for this task and underscores the necessity of the hierarchical agentic reasoning adopted in this paper.

Decision Rules

The six decision rules for hierarchical tariff rules application
The following six rules must be applied progressively from Rule 1 to Rule 6, without skipping.
Rule 1: Priority of Headings and Notes
The classification of goods shall be determined primarily according to the terms of the headings (4/6-digit HS codes) and any related Notes. Subsequent rules shall only be applied if the terms of the headings and the Notes do not suffice for classification.
Rule 2: Incomplete/Unfinished Articles and Extension to Materials/Substances
Rule 2(a): An incomplete or unfinished article (e.g., a bicycle missing wheels), if it has the essential character of the complete article, is to be classified as the complete article.
Rule 2(b): An article consisting of a certain material or substance, which retains its original character after the addition of other materials/substances (e.g., a plastic cup with a metal base), is to be classified according to the original material.
Example: An unassembled computer motherboard (which already has the function of a motherboard) is classified under heading 8473 (parts of computers).
Rule 3: Decision Logic for Goods Classifiable Under Multiple Headings
When goods are classifiable under two or more headings, classification shall be effected as follows, in order of priority:
Specificity (The more specific description shall be preferred to a more general description); Essential Character (Determined by the main material, function, or use of the goods); Last in Numerical Order (If classification cannot be determined otherwise, classify under the heading which occurs last in numerical order).
Example: An electric toothbrush (which has the attributes of both a "household appliance" and an "oral hygiene tool"): Specificity: Classified as a "domestic electro-mechanical appliance" (heading 8509) rather than a "toothbrush" (heading 9603).
Rule 4: Principle of Closest Analogy
When goods cannot be classified by applying the preceding three rules, they shall be classified under the heading appropriate to the goods to which they are most similar.
Example: Imitation leather made from a new material (not listed in the HS) is classified as "artificial leather" (heading 3921).
Rule 5: Packing Materials and Containers
Rule 5(a): Packing materials/containers presented with the goods (e.g., a jewelry box), if normally sold with the goods, are classified with the goods; otherwise, they are classified separately.
Rule 5(b): Reusable packing containers (e.g., metal gas cylinders) are classified separately.
Example: A glass bottle presented with perfume is classified under the heading for perfume (3303); however, a glass bottle sold separately is classified under 7010.
Rule 6: Hierarchical Classification at the Subheading Level
The classification of goods in the subheadings (6-digit and subsequent HS codes) of a heading shall be determined level by level, first determining the 1-dash subheading (5-6 digits), and then successively the lower-level subheadings. At each level, classification must take into account any Subheading Notes and the relationship between subheadings at the same level.
Example: After classifying goods under heading 6205 (men's shirts), the subheading is chosen based on material (cotton, man-made fibers, etc.): 620520 (of cotton) or 620530 (of man-made fibres).

Figure 12. One example of expert-written decision rules.

J Case study of Failure Modes

We identify six critical failure modes of open-source and closed-source agent systems in HSCODECOMP: (1) **Premature Decisions**: Agents commit to incorrect classification paths without collecting sufficient evidence (Figure 14 and Table 14-Grok DeepSearch); (2) **Information Misprocessing**: Agents overlook or misinterpret key product details, indicating challenges with long-context processing (Table 13 and Figure 15); (3) **Unnecessary Self-Correction**: Agents sometimes predict correct HSCodes initially but revise them incorrectly through excessive critique (Table 13, Gemini Deep Research); (4) **Reasoning Hallucination**: Agents generate plausible but factually incorrect reasoning steps (Table 13, Grok DeepSearch); (5) **Wrong Rule Application**: Models frequently miss or misuse relevant tariff rules due to their ambiguous descriptions that confuse the reasoning process, resulting in incorrect classification decisions (Figure 17); and (6) **Lack of Domain Knowledge**: Models exhibit errors due to insufficient domain-specific knowledge, such as misidentify silicone products as rubber instead of plastic (Figure 16). These limitations highlight that HSCODECOMP remains challenging for advanced closed-source and open-source systems.

Moreover, we also provide one case in Figure 13, demonstrating cutting-edge agent's failure on using human-written hierarchical decision rules. This example demonstrates how explicit rules can encourage mechanical pattern-matching over substantive functional analysis—a critical failure mode in hierarchical rule application. Specifically, the rules-based approach mechanically applied Rule 1 literal description, classifying the collar under heading 8526 (radio navigational equipment) because it contains a GPS receiver. It cited ruling NY N006896 for "GPS pet locator devices", treating GPS presence as determinative of classification. However, this misapplies Rule 3 Essential Character principle. The product's essential function is not navigation but electronic pet containment and training. GPS merely serves as the technical means to enforce boundaries. The non-rules version correctly recognized this functional distinction, classifying it under heading 8543 (electrical machines with individual functions).

K Case Study of Overthinking

Figure 18 and Figure 19 exhibit the overthinking case study of WebSailor framework.

L Case Study of Multi-modal Information

Table 15 presents experimental case studies comparing SmolAgent (GPT-5) performance with and without product image processing.

Failure of Using Human-written Decision Rules (Task#31)

Task

Product Title:GPS-enabled dog fence collar that triggers alerts when pets cross virtual boundaries.

Ground-truth HSCode: 8526910040
SmolAgent + GPT-5 (without rules): Correctly predicted 8526910040 (Rule 1)

SmolAgent + GPT-5 (with rules): Incorrectly classified as 8543709860 (Rule 3)
Used Rule 1 and 3 can be found in Figure I.1

Failure Analysis:
The rules-based approach mechanically applied Rule 1's literal description, classifying the collar under heading 8526 (radio navigational equipment) because it contains a GPS receiver. It cited ruling NY N006896 for "GPS pet locator devices", treating GPS presence as determinative of classification. However, this misapplies Rule 3's Essential Character principle. The product's essential function is not navigation but electronic pet containment and training. GPS merely serves as the technical means to enforce boundaries. The non-rules version correctly recognized this functional distinction, classifying it under heading 8543 (electrical machines with individual functions).

Figure 13. The case of failure of agent on applying hierarchical human-written rules.

Error: Premature Decisions

Task

Product Title: PEN-F PU Leather Half Case for Olympus PEN-F Digital PENF Camera Brown/Black/Coffee

Product Attributes:

```
Origin: Mainland China                          Measurement unit: 100000015
Use: Mirrorless System Camera                   Package size - length (cm): 15
Material: PU                                    Model Number: PEN-F
Brand Name: NiYi                                Type: Camera Bags, Hard Bag
each pack: 1                                    Package size - width (cm): 8
Package weight: 0.200                           Style: handbags
Package: No                                     Package size - height (cm): 5
```

Product Price: 9.8 USD
Category: Consumer Electronics → Accessories & Parts → Camera Bags & Cases

Search Query (Premature Focus)

Search query: `HTSUS camera case` **`4202.92`** `United States 10-digit code camera cases outer surface of plastic sheeting.`

Note: The agent immediately narrowed the search to Heading 4202 (Trunks, Suitcases, etc.) based on the keyword "Case", ignoring the specific material implications of "PU Leather" (Plastics).

HSCode Description

Wrong Direction (Heading 4202):
42 Articles of leather; saddlery and harness...
4202 Trunks, suit-cases, vanity-cases... camera cases...:
4202.92 ... (Incorrect for PU/Plastic composition in this context)

True Label (Chapter 39):
39 Plastics and articles thereof
3926 Other articles of plastics...:
392690 Other:
39269099 Other
3926909989 Other

Analysis

The accurate code is **3926.90.9989**. This first search query leads to a wrong direction since it decides the item is under **4202** and ignored it is **PU leather** which is chemically **plastic**. The agent committed a "Premature Decision" error by locking onto the "Camera Case" function (Heading 4202) before verifying if the material composition (PU) excluded it from that chapter, and later turns did not realize this ignorance.

Figure 14. Early wrong search query leads to wrong result.

Error: Real-world noise

Task

Product Title: 10/30/60PCS INS Blue Color PVC Sticker Aesthetic Hand Accounting DIY Decoration Scrapbooking Korean Stationery Supplies

Product Attributes:

```
Origin : Mainland China
Size : M
Material : Plastic
LOT_NUM : 1
PRODUCT_UNIT : 100000015
PACKAGE_TYPE : sell_by_piece
Package weight : 0.030
Age : >3 years old
Shape : malformed
Package size - length (cm) : 20
Model Number : sticky PVC stickers
Hign-concerned Chemical : None
Brand Name : NoEnName_Null
Package size - width (cm) : 15
Package size - height (cm) : 3
```

Product Price: 0.77 USD
Category: Office & School Supplies → Stationery Sticker → Stationery Sticker

Search query and Prediction

Search query: `site:rulings.cbp.gov PVC stickers classification 4911.91`

Prediction: 4911.91.4040

Classification path: Chapter 49 → 4911 → 4911.91 → 4911.91.4040

Decision rationale (Excerpt): The product consists of decorative, self-adhesive PVC stickers... CBP consistently classifies decorative stickers ... under 4911.91 when the printing is the primary purpose...

HSCode Description

Predicted (Wrong):
49 Printed books...
4911 Other printed matter...
491191 Pictures, designs...
4911914040 Other

True Label:
39 Plastics and articles thereof
3919 Selfadhesive plates...
391910 In rolls...
3919102055 Other

Analysis

The accurate code is 3919.10.2055. The material is PVC plastic... The misclassification error stems from **real-world noise** in the product title. This noise takes the form of subjective marketing terms (e.g., **Aesthetic**, **INS**, **Korean Stationery**) and application-focused words... This noise creates a misleading narrative that the printed design is the product's essential character.

Figure 15. Real-world noise forms and analysis.

Error: Lack domain knowledge

Task

Product Title: 20/50/100PCS Pink Nitrile Gloves Disposable Latex Powder Free Gloves for Household Cleaning Beauty Salon Working Kitchen Gloves

Product Attributes:

```
Origin : Mainland China                    Brand Name : INMAICH
Type : Disposable                          each pack : 1
Measurement unit : 100000015               Use : Cleaning
Classification : Oversleeve                Style : Brief
Material : PVC                             Model Number : Nitrile Disposable Gloves
Sold in : sellbypiece
```

Product Price: 11.56 USD
Category: Home & Garden → Household Merchandises → Disposable Household Cleaning Products

Search query and Prediction

Search query: `HTSUS nitrile disposable gloves HS code 10-digit 4015.19 CBP ruling`

Prediction: 4015.19.1010

Decision rationale (Excerpt): Material and nature: The product is described and marketed as "nitrile gloves."... This places the goods within Chapter 40 (rubber), not plastics...

Notes: If the goods were actually made of PVC/vinyl (plastic) rather than nitrile rubber, they would instead fall in Chapter 39... However, based on the stated product title... the correct US HTSUS is 4015.19.1010.

HSCode Description

Predicted (Wrong & Obsolete):
40 Rubber and articles thereof
4015 Articles of apparel... of vulcanized rubber:
- Gloves, mittens and mitts:
401519 Other:
40151911 Seamless
4015191110 Disposable (Obsolete)

True Label:
39 Plastics and articles thereof
3926 Other articles of plastics...:
392620 Articles of apparel...:
39262010 Seamless
3926201020 Disposable

Analysis

This failure is a classic example of **lacking critical domain knowledge**, demonstrated by an inability to resolve **contradictory product data** and the use of an **outdated HTS code**. The analysis was fundamentally flawed because it prioritized the marketing-oriented *Product Title (Nitrile Gloves)* over the explicit technical attribute, *Material: PVC*. This **critical error** led the analyst to misclassify the item as a **rubber** product under **Chapter 40**. A domain expert would recognize that a **specific material attribute is a more reliable basis** for classification... Compounding this mistake, the predicted code, `4015.19.1010`, is an **obsolete tariff number**...

Figure 16. Lack of domain knowledge: Conflict between title and attributes, and use of obsolete code.

Error: Wrong rule application

Task

Product Title: Fashion TYPE 2 DIABETIC Bracelet Soft Silicone Bracelet With Type 2 Diabetes Medical Warning Bracelet Hot Sale

Product Attributes:

```
Origin : Mainland China
Measurement unit : 100000015
Fine or Fashion : Fashion
Gender : Unisex
Item Type : BRACELETS
Metals Type : None
Brand Name : NPKDS
each pack : 1
Package weight : 0.013
Setting Type : None
Compatibility : All Compatible
Sold in : sell_by_piece
Chain Type : Rope Chain
Shape pattern : Round
Package size - length (cm) : 15
Material : Silicone
CN : Fujian
Model Number : SL-268
Hign-concerned Chemical : None
Function : None
Package size - width (cm) : 15
Style : TRENDY
Clasp Type : None
Package size - height (cm) : 2
Bracelets Type : Hologram Bracelets
```

Product Price: 1.71 USD
Category: Jewelry & Accessories → Fashion Jewelry → Bracelets

Search query and Prediction

Search query: `US HTS 7117.90 silicone bracelet imitation jewelry HS code medical alert bracelet`

Prediction: 7117.90.7500

Decision rationale (Excerpt): Product is a bracelet made of silicone (a plastic), marketed as a fashion/medical warning bracelet... It is not of precious metal... functions as personal adornment... fitting the terms of "imitation jewelry" of heading 7117...

HSCode Description

Predicted (Wrong):
71 Natural or cultured pearls... imitation jewellery...
7117 Imitation jewelry:
711790 Other:
7117907500 Of plastics

True Label:
39 Plastics and articles thereof
3926 Other articles of plastics...:
392690 Other:
39269099 Other
3926909989 Other

Analysis

This failure is a classic case of **Wrong Rule Application**, where a basic rule was incorrectly prioritized over a decisive, higher-level one. The agent mistakenly applied the simple rule of **classification by apparent form**, seeing a *bracelet* and defaulting to the seemingly specific **Chapter 71 (Imitation Jewelry)**. However, the correct, overriding rule that should have been applied is the Rule of **Essential Character**, which embodies the important principle of **Function Over Form** in customs classification. In this case, the text **"TYPE 2 DIABETIC"** was the key that defined its core **function**, fundamentally changing the product's essence from an ornament to a **utilitarian medical alert device**. An article with such an overriding utilitarian function is legally excluded from the scope of **Chapter 71**. Therefore, the correct rule application mandates that the item be classified by its **function** and material under **Chapter 39**, not by its misleading physical form.

Figure 17. Wrong Rule Application: Confusion between Imitation Jewelry and Functional Articles.

Table 13. Comparative Analysis of HSCode Prediction for a Forklift Safety Cage mentioned above.

Analysis Dimension	**SmolAgent**	**Gemini DeepResearch**	**Grok DeepSearch**	**Manus**
Final HTSUS Code	8431.20.0000 (Correct)	7326.90.8688 (Incorrect)	8428.90.0290 (Incorrect)	8427.90.0020 (Incorrect)
Core Logic Explained	Based on the core principle of HTSUS Section XVI, Note 2, the cage is a **part** as it is 'solely or principally for use with' a forklift (heading 8427). Its design, function, and identity are entirely dependent on the forklift.	The argument is based on the **'part vs. accessory' distinction**. It posits the cage is not an **'indispensable' part**, but an optional 'accessory'. Since accessories are precluded from 8431, classification defaults to its constituent material (steel).	Characterizes the cage as a **functional piece of machinery**. The rationale is that it enables a new function and **incorrectly compares it to complex attachments** with their own mechanics (e.g., rotators, clamps).	Characterizes the cage as a **complete 'aerial work platform'**. The core argument is that its **'4 universal wheels' constitute a 'mobile base'** per the legal notes, thus assembling it into a complete vehicle.
Key Flaw Analysis	This approach correctly identifies the product's primary use and applies the controlling legal note directly, which is the standard and most reliable method for classification.	This is overthinking because the model became **fixated on a complex, secondary legal nuance** (part vs. accessory) while **ignoring the more direct, primary rule** ('solely or principally for use with'), leading to an unnecessarily complicated and incorrect conclusion.	This is an analysis hallucination because the model **invents characteristics the product lacks**, effectively **treating a passive structure as an active machine**. The entire analysis is built on this fabricated, non-existent product feature.	The model misses key product information by **misunderstanding the function of a key feature** (the wheels). It correctly identifies the wheels but misses their trivial context (ground convenience), instead **mistaking them for a vehicle's chassis**, which invalidates the entire classification.
Failure Modes	**Correct**	**Unnecessary Self-Correction**	**Reasoning Hallucination**	**Information Misprocessing**

Table 14. Comparative Analysis of AI Model Classifications for a Cylinder Shock Absorber mentioned above.

Analysis Dimension	**SmolAgent**	**Gemini DeepResearch**	**Grok DeepSearch**	**Manus**
Final HTSUS Code	8487.90.0080 (Correct)	8412.21.0075 (Incorrect)	8302.49.6085 (Incorrect)	8412.31.0080 (Incorrect)
Core Logic Explained	Based on the hierarchical structure of HTSUS Chapter 84, the shock absorber is a **generic machinery part**. After systematically eliminating more specific headings, it correctly classifies the item in the **residual heading 84.87** for parts "not elsewhere specified."	Characterizes the product as an **active hydraulic motor**. The logic is that because it is a linear-acting hydraulic device, it must be a "motor" under heading 84.12, which is an apparatus that **generates force or motion**.	It correctly identifies the passive function but then classifies it as a **simple base metal fitting**. The argument is that since it's not a motor, its classification defaults to a general heading for **common hardware and accessories**.	Characterizes the product as an **active pneumatic motor**. The rationale is based on the keyword "Pneumatic Cylinder" in the title, concluding it must be an **actuator that performs work** under heading 84.12.
Key Flaw Analysis	This approach correctly identifies the product's non-specific nature and applies the HTSUS's hierarchical structure and residual headings, which is the standard and most reliable method for such goods.	This is a fundamental misunderstanding of the product's function, as the model **mistakes a passive energy-dissipating device (a damper) for an active power-generating device (a motor)**. It confuses braking with accelerating.	The model makes a decision with insufficient information about HTSUS structure because it **fails to consider the critical distinction between Chapter 83 (simple fittings) and Chapter 84 (machinery)** and thus **underestimates the product's nature** as a piece of machinery.	This is a fundamental misunderstanding of the product's function, as the model is **misled by an inaccurate keyword** and **mistakes a passive damper for an active motor**, ignoring contradictory product attributes.
Failure Modes	**Correct**	**Information Misprocessing**	**Premature Decisions**	**Information Misprocessing**

Case 1: WebSailor Overthink.

Task

Product Title: Men's and Women's Fashion Printed T-shirts, Men's and Women's Fashion Shirts, WWE John Scene, Winning This Day, Summer, New

Product Attributes:

```
Origin : Mainland China
Gender : Men
Hign-concerned Chemical : None
Applicable Scene : Daily
each pack : 1
Style : Casual
Package size - height (cm) : 10
Measurement unit : 100000015
Material : POLYESTER
Craft of Weaving : Knit
Brand Name : NoEnNameNull
Product Care Instructions : Machine Wash
Pattern Type : 3D
Closure Type : Pull On
Package size - length (cm) : 10
Fabric Type : Broadcloth
Special Features : Breathable,Lightweight
Place Of Origin : China (mainland)
Applicable Season : Spring And Summer
Sleeve Length(cm) : short (4-16inch)
Collar : O-Neck
Tops Type : TEES
Item Type : tops
CN : Jiangxi
Package size - width (cm) : 10
Hooded : No
Package weight : 0.200
Sold in : sellbypiece
```

Product Price: 74.96 CNY
Category: Men's Clothing → Tops & Tees → T-Shirts

First step

Websailor:
*We need the precise 10-digit HTSUS classification for a men's knit T-shirt of man-made fiber (polyester)... The common 10-digit code used by importers ... is **6109.90.1050 (men's or boys')**. I will confirm using CBP rulings...*

`name : search , arguments : { query : [ HTSUS 6109.90.1050 men t-shirt man-made fibers ... ]}`

Smolagents: search query: 'HTSUS 6109.90 T-shirts of man-made fibers 10-digit statistical suffix men

HSCode Description

Websailor wrong prediction:
61 Articles of apparel...
6109 Tshirts, singlets...
610990 Of other textile materials:
61099010 Of manmade fibers
6109901050 Women's (639)

Smolagents correct prediction:
Section XI Textiles...
61 Articles of apparel...
6109 Tshirts...
610990 Of other textile materials:
61099010 Of manmade fibers
6109901007 Men's (638)

Figure 18. Case study comparing Websailor and Smolagents reasoning on HS Code classification.

Case 2: WebSailor Overthink and Directly Answer.

Task

Product Title: SPRAY GUN with adapter 400cc Mix tank 1.3/1.8mm nozzle air spray gun with quick-connect wash-free tank paint Mixing Cup sprayer

Product Attributes:

```
Origin : Mainland China
Measurement unit : 100000015
Use : Commercial Manufacture
Package size - length (cm) : 30
Model Number : GTI
Feed Type : Gravity
Brand Name : NTOOLS
each pack : 1
Package weight : 1.300
Application : Paint Spray Gun
Sold in : sellbypiece
Power Type : Pneumatic
Battery Included : No
Cup Volume : 400cc
Nozzle Diameter : 1.3mm
Hign-concerned Chemical : None
Certification : CE
Type : HVLP
Package size - width (cm) : 22
No-Load Speed : 1200rpm
Package size - height (cm) : 20
```

Product Price: 98.0 USD
Category: Tools → Power Tools → Spray Gun

First step

Websailor:
We need the precise 10-digit US HS (HTSUS) code for a pneumatic paint spray gun... Specifically, subheading 8424.20 covers "Spray guns and similar appliances."... The 10-digit breakouts... fall in 8424.20.10 at the 8-digit level... Therefore, the full 10-digit code is ***8424.20.1000*** *or* ***8424.20.1090****... Now present classification path...* 8424.20.1000
...

SmolAgents:
search query: HTS 842420 paint spray guns 10 digit statistical suffix

HSCode Description

Websailor wrong prediction:
84 Nuclear reactors...
8424 Mechanical appliances...
842420 Spray guns and similar appliances:
8424201000 Simple piston pump sprays and powder bellows

Smolagents correct prediction:
84 Nuclear reactors...
8424 Mechanical appliances...
842420 Spray guns and similar appliances:
8424209000 Other

Figure 19. Comparison of reasoning paths for Spray Gun classification (Case 2).

Table 15. Vision vs Non-vision: Four Representative Cases with Key Evidence and Reasons

Image	Product Information	Non-vision HS (10-digit) + brief rationale	With-vision HS (10-digit) + brief rationale	Vision advantage (reason category)
	Title: Royal Wedding Bouquet Rhinestone Bride and Bridesmaid Hand Flowers Handmade Bridal Bouquet. **Category:** Artificial Decorations. **Attributes:** Material: **Silk (natural textile)**, not plastics; bouquet of artificial flowers with rhinestones.	**6702.90.3500** (Artificial flowers of *man-made fibers*). Assumes "silk flowers" = polyester; treats bouquet as MMF artificial flowers.	**6702.90.6500** (Artificial flowers, *other than man-made fibers*). Images/text indicate **silk fabric (natural fiber)**; rhinestones do not change essential character (GRI 3(b)).	**Material/fiber identification** (natural silk vs assumed polyester).
	Title: OTF knife parts, aviation aluminum tactical handle kit. **Category:** Hand Tools / Knife accessories. **Attributes:** **Handle/-parts only**, no blade present; aluminum body with clip, actuator, screws.	**8211.93.0035** (Folding/pocket knives). Interprets listing as a *complete folding knife* rather than components.	**8211.95.9000** (Handles of base metal, other). Visuals confirm **no blade**; essential character is the base-metal handle assembly (parts provision applies).	**Completeness/parts identification** (parts-only vs whole article).
	Title: Acrylic buckle, beaded/openwork shoulder bag with inner pouch. **Category:** Women's Handbags. **Attributes:** Outer surface is **beads/openwork lattice (ABS/acrylic)**, not plastic sheeting; linen pouch is interior.	**4202.22.1500** (Handbags with outer surface of *sheeting of plastics*). Assumes exterior is plastic sheeting based on "ABS."	**4202.29.9000** (Other). Images show **beads/rings openwork**, not "sheeting of plastics"; classification by outer surface (Additional U.S. Note 2 to Ch. 42).	**Structural outer-surface feature** (beads/openwork vs plastic sheeting).
	Title: The Simpsons character collectible paper cards. **Category:** Printed matter / collectibles. **Attributes:** Cards bear **pictures only**; no rules-based deck specified; **no lithographic process evidence**.	**4911.99.6000** (Other printed matter, often linked to lithographic printing). Assumes lithographic process without explicit evidence.	**4911.91.4040** (Pictures, designs and photographs; other). Visuals confirm **picture-only cards** and lack of lithographic evidence/criteria; not a playing-card game.	**Process/evidence-based exclusion** (no lithography evidence; pictures-only).