Right Code, Wrong Reasons: An Audit Trail Benchmark for Customs Classification

Anonymous ACL submission

Abstract

Several high-stakes administrative tasks require supporting verdicts and decisions with auditable reasoning. While the application of LLMs in this regard has been studied, existing benchmarks are unsuitable as they either score only the final answer or evaluate citation quality separately from outcome correctness. We introduce **HS-BENCH**, a benchmark for customs classification, to study this problem. The task involves assigning tariff codes to products with reasoning grounded in legal authorities. HS-BENCH consists of 1,390 product descriptions paired with supporting legal authorities across diverse sources such as dictionary definitions, statutory notes, and prior rulings. We propose using Strict Support Accuracy (SSA), a metric that credits a prediction only when both the outcome and the reasoning is well supported. Evaluating frontier models in closed-book and oracle-evidence regimes, we find that classification accuracy and SSA diverge sharply: models reach the correct outcome through incorrect reasoning. We release the dataset, extraction pipeline, and evaluation code to support work on auditable LLM systems in administrative law.

1 Introduction

LLMs are increasingly studied for high-stakes administrative tasks — taxation (Blair-Stanek et al., 2026), public services (Moon et al., 2026), regulatory compliance (Gui et al., 2025), and tariff classification (Yang et al., 2025). In these settings, correct outcomes must be supported by auditable reasoning: tax determinations must cite the controlling statute and regulation, tariff rulings must invoke the correct GRI and applicable notes, and regulatory decisions must point to the specific provision applied. A correct outcome supported by the wrong authority is operationally indistinguishable from a wrong outcome under review, making joint evaluation of outcome and reasoning essential.

Existing benchmarks do not evaluate these properties jointly. Early legal-NLP benchmarks focused on classification tasks without reasoning or grounding (Chalkidis et al., 2022; Niklaus et al., 2023). LegalBench (Guha et al., 2023) introduced IRAC-style reasoning tasks but did not require evidence attribution, while LawBench (Fei et al., 2024) and LexEval (Li et al., 2024) rely on overlap-based generation metrics such as ROUGE. More recent exam-oriented benchmarks attempt to evaluate legal reasoning more directly, but differ substantially in methodology. GreekBarBench (Chlapanis et al., 2025) and KCL-Essay (Oh et al., 2026) use expert-validated rubrics and provide question-aligned authorities that should be cited, whereas OAB-Bench (Pires et al., 2026) and LEXam (Fan et al., 2026) rely on LLM judges without external authorities. However, because these benchmarks are grounded in pedagogical exam evaluation, citation correctness is not evaluated jointly with the final outcome. Exams naturally permit partial credit for incomplete or weakly supported answers, making them poor proxies for workflows where procedural validity depends on both the decision and the supporting authority.

The gap shows up most acutely in customs, where ATLAS (Yuvraj and Devarakonda, 2025) and HSCodeComp (Yang et al., 2025) score hierarchical HS-code accuracy but evaluate neither reasoning nor evidence attribution; a model with high accuracy but fabricated grounding is not deployable in the workflows that motivated the benchmark.

Customs is auditable-by-design. Unlike tax, medical coding, or regulatory compliance, where precedent is often implicit, the General Rules of Interpretation (GRIs), statutory notes, doctrinal tests, and prior CBP rulings are cited inline in every decision — the cleanest substrate for testing whether LLMs can match the legal path a practitioner would follow.

Adjacent work has begun formalizing this dis-

tinction between answer correctness and attribution correctness. In document QA, Ma et al. (2026) show that frontier models often answer correctly while citing incorrect evidence regions ("attribution hallucination"), and propose Strict Attributed Accuracy (SAA), which requires both answer and citation correctness. We bring this lens to customs classification with **HS-BENCH**, a benchmark of 1,390 CBP product-classification rulings paired with the supporting legal authorities underlying each decision, extracted through a two-stage pipeline validated by legal experts. We further introduce **Strict Support Accuracy (SSA)**, a support-faithful metric that credits predictions only when both the tariff code and supporting legal basis are correct, to evaluate the task in our benchmark.

Across four frontier models evaluated in closed-book, retrieval, and oracle settings, we find a persistent gap between answer accuracy and support-faithful reasoning. Even with oracle access to supporting authorities, frontier models achieve nearly 94% tariff-line accuracy while passing SSA on only $\sim 60\%$ of examples, revealing a deployment-critical failure mode hidden by conventional accuracy metrics.

Contributions. To summarize, our work introduces **HS-BENCH**, a benchmark for evaluating support-faithful customs classification; proposes **Strict Support Accuracy (SSA)**, a metric for jointly measuring outcome correctness and legal grounding; and provides the first systematic analysis of the gap between classification accuracy and procedural validity across frontier LLMs.

2 HS-Bench

The CBP releases rulings and decisions on applications by importers through the Customs Rulings Online Search System (CROSS). We curate the dataset to include only rulings from January 2024 to April 2026 to reflect the recent legal landscape. Rulings related to classification under HS which are identified by each ruling's subject category.HS-BENCH consists of triples of product descriptions, their assigned classification and the reasoning for the classification. We describe the 2-stage process for extracting these triples from the rulings in this section.

Product Extraction. In the first stage, product descriptions and their assigned codes are extracted from 2,282 rulings. A single application can request tariff rulings for multiple products. The product descriptions are a synthesized combination of facts submitted by the applicant and assertions by the department based on physical inspection and laboratory tests. This mirrors tariff classification in practice, where establishing the product's factual profile is separate from classification reasoning. We filter out rulings which do not assign a final classification.

Product description
A foldable rolling walker (rollator) with a lightweight aluminum frame, four wheels, brakes, height-adjustable handles, and a seat with backrest; designed to support persons who have difficulty walking.
Ground truth answer
Heading 8713 ("carriages for disabled persons") is excluded because a rollator is not a "carriage" in the common and commercial meaning of that term, as established in **Toyota Motor Sales v. United States (1984)** and **United States v. Quon Quon Co. (1959)**, and consistent with the **Oxford English Dictionary**. The product is therefore classified as an orthopedic appliance under **9021.10.0090**.
Predicted answer
A rollator is not a "carriage," so heading 8713 does not apply. The product is an orthopedic appliance, classified under **9021.10.0090**.

Table 1: Representative HS-BENCH instance. The predicted answer reaches the same 8-digit code as the ground truth but omits the two federal-court precedents and dictionary definition that justify excluding heading 8713.

Reasoning Extraction. In the second stage, the reasoning used for each product's classification is extracted along with the set of grounding authorities cited. Every reasoning trace requires engagement with the GRIs and the chosen tariff code. A subset also requires citing from a set of possible authorities as described in Table 2. Following the methodology of Oh et al. (2026), where only questions that reference at least one citation are included, traces with sparse reasoning traces are excluded. These are traces which do not invoke GRIs explicitly and do not engage with any external authorities or competing headings. More details about the dataset can be found in Appendix B.

The final dataset consists of 1,390 product descriptions and their respective classification and reasoning aligned to 7,284 grounding authorities.

Validation. We use Kimi K2.6 (Moonshot AI, 2026) for both stages of extraction. We optimize

Authority	Description	N	%
Tariff entries	Competing tariff entries from the HTSUS distinguished in the analysis	430	42.5
HTSUS notes	Statutory notes in the HTSUS (chapter and section notes)	242	39.4
Explanatory notes	WCO commentary on the scope of each heading	307	38.3
CBP rulings	Prior rulings issued by the CBP	695	27.3
Case laws	Judicial precedent from U.S. federal courts	153	13.6
Definitions	Common and technical dictionary definitions	64	4.7

Table 2: Authorities cited in HS-Bench reasoning traces. N is the number of unique sources of each type; % is the share of products citing at least one source of that type. See Appendix B for more details

prompts using GEPA (Agrawal et al., 2026) based on a set of 100 outputs generated using Claude Opus 4.7 (Anthropic, 2026) from a held-out set of rulings, reviewed by a legal expert. The outputs of the pipeline on 100 rulings from the dataset were reviewed at both stages before processing the complete dataset. Further details are present in Appendix B

3 Experimental Setup

Task. Given a product description d and a (possibly empty) set g of supporting authorities, the system produces (i) an HTSUS classification $c \in \mathcal{C}$ at the 10-digit statistical reporting level and (ii) a reasoning trace r that may invoke GRIs, section and chapter notes, Explanatory Notes, doctrinal tests, and competing headings to justify c. The trace is parsed into atomic claim units using the same decomposition procedure applied to the gold reasoning.

Evidence Access. We evaluate two evidence-access regimes. In *Closed-book* the input is the product description alone, bounding parametric knowledge. In *Oracle* we assume a perfect retriever and supply the full bodies of the authorities cited by the bench, additionally revealing the gold 6-digit subheading with all of its 8- and 10-digit children - the strict evidence-usage ceiling.

Retrieval. The oracle regime represents the ceiling performance that evidence access can deliver when both the gold authorities and the gold subtree are known. To highlight what a realistic deployed system reaches without either, we evaluate two retrieval-augmented conditions in which both supporting authorities *and* candidate tariff headings are retrieved rather than supplied. Each retriever runs two passes. *FT-dense* uses two fine-tuned dense retrievers - one over product-description-to-ruling pairs and one over description-to-heading pairs. The pairs are constructed from CROSS rulings outside HS-BENCH
Graph walks two graphs of the CBP corpus. Candidate rulings are retrieved from a homogeneous graph with rulings as nodes, with edges between rulings citing each other. Candidate tariff entries are retrieved from a heterogeneous graph of rulings and headings as nodes, with additional edges between rulings and their references headings.

Metrics. *Heading accuracy* is exact match on the first k digits of the predicted code against the gold code. We report it for $k = 8$, which represents the classification that determines the duty rate. For reasoning evaluation, a two-pass LLM judge first decomposes the gold reasoning into atomic claim units (Liu et al., 2023) and then returns a binary entailment label per claim against the candidate audit trail; per-product coverage $\mathrm{cov}(x_i)$ is the fraction of gold claims marked entailed (recall over decomposed gold claims, not retrieval recall; we use recall rather than F1 because the gold reasoning is terse while candidate audit trails are often more thorough). *Strict Support Accuracy at k digits* (SSA_k) credits a prediction only when the code is correct at k digits *and* $\mathrm{cov}(x_i) \geq \tau = 0.8$:

$$\mathrm{SSA}_k = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\mathrm{code}_k(x_i)] \cdot \mathbb{1}[\mathrm{cov}(x_i) \geq \tau]. \tag{1}$$

This is a simpler joint structure than CiteVQA's SAA (Ma et al., 2026), which conjoins answer correctness with a per-citation Relevance-or-Recall disjunction; the discrete-authority setting of customs permits a single coverage primitive over per-product gold claims. The headline metric is SSA_8. The judge is calibrated and debiased using prediction-powered inference against three lawyer-annotators on a stratified 96-product sample (Appendix C).

Models. We evaluate four frontier models spanning the closed- and open-weights frontier: Claude Opus 4.7 (Anthropic, 2026) and GPT-5.4 (closed), DeepSeek v4-pro and Kimi K2.6 (Moonshot AI, 2026) (open). Reasoning is disabled across all

Model	Regime	Acc@8	SSA@8
Oracle ladder			
Opus 4.7	Closed-book	0.458	0.273
	Oracle	**0.935**	**0.602**
GPT-5.4	Closed-book	0.459	0.241
	Oracle	**0.937**	**0.643**
DeepSeek V4-Pro	Closed-book	0.477	0.190
	Oracle	0.844	0.454
Kimi K2.6	Closed-book	0.294	0.100
	Oracle	0.828	0.454
Retrieval			
GPT-5.4	FT-dense	0.730	0.410
	Graph	0.637	0.390
DeepSeek V4-Pro	FT-dense	0.604	0.359
	Graph	0.635	0.361

Table 3: Heading accuracy (Acc@8) and Strict Support Accuracy (SSA@8) across evidence regimes — **bold** marks the headline cells. Debiaising results, 95% confidence intervals and full digit ladder in Appendix D.

four; the output schema and decoding parameters are held fixed across cells, with the schema constrained so audit-trail parsing is deterministic.

4 Analysis

Right code, wrong reasoning. Table 3 reports Acc@8 and SSA@8 across the four frontier models and two evidence-access regimes. Under full Oracle access — the gold 6-digit subheading leaked and the cited authorities supplied as full text — the two closed-frontier models near-saturate tariff-line accuracy (Opus 0.935, GPT 0.937) yet pass SSA on only ∼60% of items (Opus 0.602, GPT 0.643), a 30-point shortfall.

The SSA shortfall decomposes into classification failure (wrong code) and attribution failure (right code, wrong reasoning). Conditioning on items where the predicted code matches the gold, the attribution recovery rate $P(\text{cov} \geq 0.8 \mid \text{code correct})$ is 64% for Opus and 69% for GPT at the closed-frontier end, 54% for DeepSeek and 55% for Kimi at the open-frontier end — 31-46% of correctly-classified items reach the right code via reasoning that does not recover the legal authorities the bench relied on. Closed-book bounds parametric knowledge and Oracle bounds evidence access; the residual at Oracle is neither — a path divergence between the model's reasoning trace and the bench's. This is the diagnostic claim of the paper, the attribution-hallucination pattern (Ma et al., 2026) now in administrative classification.

Retrieval does not close the gap. Retrieval narrows the closed-book-to-Oracle SSA gap by 42-65% across the two frontier models with retrieval data (Table 3); the two variants (FT-dense and Graph) land within ∼2 SSA points of each other in both, suggesting the bottleneck is not retriever choice. The residual gap reflects attribution failure at Oracle, not retrieval failure - retrieval improvements alone cannot close it.

Robust across the frontier. All four frontier models show the Acc@8 / SSA@8 divergence under Oracle, with closed-frontier (Opus, GPT-5.4) recovering more of the gold reasoning than open-frontier (DeepSeek, Kimi); no model saturates. The divergence is not a metric artifact: the reasoning judge agrees substantially with three lawyer-annotators on a stratified 96-product calibration sample, with Fleiss' $\kappa = 0.64$ and pairwise agreement in the range $[0.616, 0.634]$ (Appendix C).

5 Conclusion

In high-stakes administrative reasoning, the answer is not only the outcome - it is the audit trail that makes the outcome reviewable. We introduced HS-BENCH, a benchmark of 1,390 product classifications from CBP rulings paired with the supporting legal authorities the bench cited, and SSA, a metric that credits a prediction only when both the assigned code and its supporting reasoning are correct. Across four frontier models under full oracle access, tariff-line accuracy reaches 0.83-0.94 but SSA reaches only 0.45-0.64. Conditional on a correct code, reasoning recovers the bench's cited authorities on only 54-69% of items - a path divergence between the model's reasoning and the bench's that retrieval improvements alone cannot close. The natural next step is to ask whether models can be trained - not just prompted - to follow the legal path; HS-BENCH provides the metric and dataset to measure progress. We release the dataset, extraction pipeline, and evaluation code to support follow-on work in tariff classification and adjacent administrative domains where correctness alone is insufficient.

Limitations

HS-BENCH focuses exclusive on U.S. jurisdiction, with data sourced only through CROSS. While the Harmonized System is used in over 200 countries, we note that the results of this work might not generalize to other jurisdictions as the cited authorities

and statutes vary by country. Further, the work only focuses on English-language product description, limiting application to multi-lingual domains.

The dataset reflects products for which importers sought administrative guidance and may underrepresent routine classifications. Finally, full SSA is a strict joint metric; partially correct reasoning paths that a human expert could repair under deployment are not credited.

The reasoning coverage is measured by an LLM judge against a single reference path; a valid alternate route to the same heading may score low if it does not recover the reference claims. We also do not study properties of hallucinations on the dataset.

The gold reasoning trace is generated per-product and scoped to single-product classification, excluding the cross-product distinguishing reasoning present in multi-product source rulings. This preserves the single-product input/output structure but means SSA does not directly evaluate cross-product reasoning. The source rulings retain this reasoning and could support a separate multi-product benchmark. The benchmark targets U.S. practice (HTSUS) and CBP rulings; transfer to other jurisdictions is not evaluated.

References

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alex Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. 2026. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *The Fourteenth International Conference on Learning Representations*.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. 2023. ppi_py: A python package for prediction-powered inference. https://github.com/aangelopoulos/ppi_py.

Anthropic. 2026. Claude opus 4.7. https://www.anthropic.com/news/claude-opus-4-7. Accessed: 2026-05-25.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2026. Can llms identify tax abuse? *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(45):38261–38269.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Yiqun T. Chen, Sizhu Lu, Sijia Li, Moran Guo, and Shengyi Li. 2026. Efficient inference for noisy LLM-as-a-Judge evaluation. *arXiv preprint arXiv:2601.05420*.

Odysseas S. Chlapanis, Dimitrios Galanis, Nikolaos Aletras, and Ion Androutsopoulos. 2025. GreekBarBench: A challenging benchmark for free-text legal reasoning and citations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25099–25119, Suzhou, China. Association for Computational Linguistics.

Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer, Daniel Brunner, Markus Leippold, Mrinmaya Sachan, Alexander Stremitzer, Christoph Engel, Elliott Ash, and Joel Niklaus. 2026. LEXam: Benchmarking legal reasoning on 340 law exams. In *The Fourteenth International Conference on Learning Representations*.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA. Association for Computational Linguistics.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: a collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Haoyang Gui, Thales Bertaglia, Taylor Annabell, Catalina Goanta, Tjomme Dooper, and Gerasimos Spanakis. 2025. Evaluating LLM-generated legal explanations for regulatory compliance in social media influencer marketing. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 157–171, Suzhou, China. Association for Computational Linguistics.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024. Lexeval: a comprehensive chinese legal benchmark for evaluating large language models. In *Proceedings of the 38th International Conference on Neural Information Processing*

Systems, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Yixin Liu, Alexander Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Towards interpretable and efficient automatic reference-based summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16360–16368, Singapore. Association for Computational Linguistics.

Dongsheng Ma, Jiayu Li, Zhengren Wang, Yijie Wang, Jiahao Kong, Weijun Zeng, Jutao Xiao, Jie Yang, Wentao Zhang, Bin Wang, and Conghui He. 2026. Citevqa: Benchmarking evidence attribution for trustworthy document intelligence. *arXiv preprint arXiv:2605.12882*.

Erina Seh-Young Moon, Matthew Tamura, Angelina Zhai, Nuzaira Habib, Behnaz Shirazi, Altaf Kassam, Devansh Saxena, and Shion Guha. 2026. The promises and perils of using llms for effective public services. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA. Association for Computing Machinery.

Moonshot AI. 2026. Kimi k2-6. https://www.kimi.com/blog/kimi-k2-6. Accessed: 2026-05-25.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. LEXTREME: A multi-lingual and multi-task benchmark for the legal domain. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3016–3054, Singapore. Association for Computational Linguistics.

Hongseok Oh, Wonseok Hwang, and Kyoung-Woon On. 2026. Korean canonical legal benchmark: Toward knowledge-independent evaluation of LLMs' legal reasoning capabilities. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 232–243, Rabat, Morocco. Association for Computational Linguistics.

Ramon Pires, Roseval Malaquias Junior, and Rodrigo Nogueira. 2026. Automatic legal writing evaluation of llms. In *Proceedings of the Twentieth International Conference on Artificial Intelligence and Law*, ICAIL '25, page 420–424, New York, NY, USA. Association for Computing Machinery.

Walter J. Rogan and Beth Gladen. 1978. Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, 107(1):71–76.

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

Pritish Yuvraj and Sivanvitha Devarakonda. 2025. ATLAS: Benchmarking and adapting large language models for global trade via harmonized tariff code classification. In *NORA: The First Workshop on Knowledge Graphs & Agentic Systems Interplay*.

Size	
Products	1,390
Unique rulings	1,355
Description length (words)	
Mean	91.5
Median	84
Std. dev.	49.2
Min / Max	6 / 514
Authorities per product	
Mean	2.70
Median	1.0
Std. dev.	3.53
Max	27

Table 4: HS-Bench dataset summary statistics.

Authority	N	Source
Tariff entries	689	Harmonized Tariff Schedule of the U.S
HTSUS notes	753	Harmonized Tariff Schedule of the U.S.
Explanatory notes	753	N/A
CBP rulings	1,033	CROSS
Case laws	436	CAP, CourtListener
Definitions	94	Wiktionary

Table 5: Authorities cited in HS-Bench reasoning traces and their sources.[1]

A Background

B Dataset Details

Table 4 reports additional details about the dataset. The dataset covers 74 of the 97 chapters of the harmonized system, with 1,083 unique 10-digit codes.

Table 5 reports additional details about the various authorities present in the dataset.

Extraction validation. We asked legal experts in the domain of customs law to annotate the outputs of the the extraction pipeline. For product description extraction, annotators were provided with the extracted description and the full text of the source ruling. They were instructed to correct instances if they were incomplete (missing information from the product description), leaky (included information from the reasoning), or incorrect (included details of other products present in the application). Both of the observed errors leaked some reasoning into the description.

For reasoning extraction, annotators were provided with the product description, extracted reasoning, and the full text of the source ruling. They were instructed to correct instances if the extracted reasoning did not reflect the reasoning used for the classification of the provided product - inclusion of non-classification reasoning claims or missing cited authorities. Both of the observed errors included some non-classification claims.

Given the low error rate of 3%, the pipeline was accepted and the remaining dataset was processed with it.

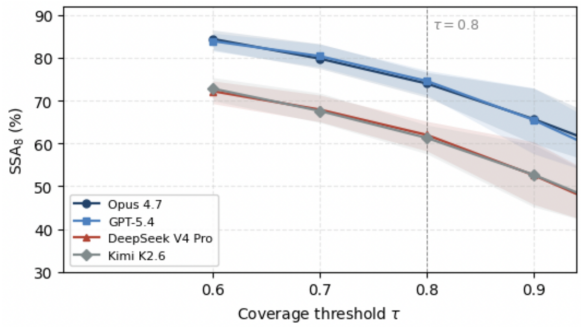

Figure 1: SSA_8 vs. pass threshold τ in the complete-oracle regime. Point estimates are PPI++ debiased; shaded bands are bootstrap 95% CIs. The headline threshold $\tau = 0.8$ is marked. Cross-model ranking is preserved at every threshold.

C Implementation Details

Judge configuration. The judge is Gemini 3.1 Flash-Lite, accessed via OpenRouter and held fixed across every cell so judge identity never co-varies with the candidate model or the evidence-access regime. It runs in two passes. The first reads the gold reasoning trace and breaks it into atomic claims, each tagged with the authority it cites; an average gold trace yields 4.5 claims. The second takes the candidate audit trail and returns, claim by claim, whether the candidate entails the claim. Both passes emit JSON; the schemas are in the supplementary material. Per-product coverage is the fraction of claims in the second pass labeled entailed.

We set the SSA pass threshold to $\tau = 0.8$. We picked this from the calibration sample: it is the lowest coverage at which all three annotators agreed a trace "substantiates the assigned code". Figure 1 sweeps $\tau \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ across the four frontier models in the complete-oracle regime; the model ranking is stable at every threshold, and the gap to plain accuracy narrows as τ falls and widens as it rises.

Judge calibration against humans. The judge is a noisy estimator of the human-majority coverage

labels SSA is defined against, so the raw judge rates are biased estimates of the quantity we care about. This section gives the human data that lets us quantify the bias and the estimator that absorbs it.

We sampled 96 products from four (eval-setting $\times$ candidate model) cells where outputs already existed: closed-book, lite-oracle, and complete-oracle, each crossed with Claude Opus and Sonnet. Within each cell we split products by judge coverage into a low-coverage bin (below 0.7) and a high-coverage bin (0.7 or above), and gave products in the borderline band $[0.3, 0.7]$ a $2\times$ inclusion weight, since this is the band where judge errors most often flip the SSA outcome at $\tau = 0.8$. Decomposing each product's gold trace and pairing every claim with the judge's verdict on the candidate trace produced $n = 382$ (judge, human) claim pairs. Per-product sampling weights are stored alongside the calibration set so the estimator can reweight back to the unstratified population.

Our legal experts trained on the SSA rubric labeled the 382 pairs as entailed or not entailed. The majority vote across the three is the human-majority label the judge is calibrated against. On the 186 claims that all three saw, Fleiss' κ was 0.64 (substantial), and all three pairwise Cohen's κ were in $[0.616, 0.634]$. Ties and malformed verdicts were dropped before pairing.

On these pairs the judge has per-claim sensitivity 0.919 and specificity 0.803. The two error rates are asymmetric: the judge misses roughly 8% of truly-entailed claims but falsely accepts roughly 20% of unentailed ones, so on aggregate it over-credits coverage. The asymmetry is what rules out a scalar mean-shift correction - the direction of bias in any given cell depends on the underlying prevalence of entailed claims there, since the two error types push the estimate in opposite directions.

PPI++ combines the 382 paired claim labels with the $\sim$6,500 judge-only labels in each cell into a single debiased mean targeting the human-majority coverage, and the asymptotic CIs propagate both eval-set sampling variance and calibration-set sampling variance. The calibration-set sampling weights are applied before debiasing. Implementation is `ppi_py` (Angelopoulos et al., 2023) at $\alpha = 0.05$. As a cross-check, the Rogan-Gladen (Rogan and Gladen, 1978) estimator on the same data agrees with PPI++ within $\sim$2 pp on per-cell point estimates but produces 20–35% wider CIs, consistent with the efficiency results of Chen et al. (2026).

Decoding setup. All four candidate models are evaluated with chain-of-thought reasoning explicitly disabled, via OpenRouter. We wanted the audit trail we score to be what the model commits to on the record (its written reasoning paragraph), not a deliberation summarized post hoc from a hidden scratchpad. Decoding parameters are held fixed across every cell: greedy / deterministic sampling, a fixed per-generation token budget, and schema-constrained output so the heading code and the reasoning paragraph are emitted as separate, deterministically parseable fields. The judge runs under the same regime: reasoning disabled, deterministic sampling, schema-constrained output for both the claim-decomposition stage and the claim-verification stage.

Sensitivity to reasoning. To probe whether disabling reasoning materially distorts the picture, we ran two contrasts. First, a paired comparison on a 50-product random sample between GPT-5.4 with reasoning off and Gemini-3.1-Pro with low-effort reasoning forced on (its endpoint requires reasoning to be enabled). On this sample, tariffline accuracy differed by less than two percentage points and the LiteOracle versus PureOracle ordering was identical. Our released pipeline exposes reasoning as a model-level switch, so any cell can be rerun with reasoning enabled (and the corresponding token-cost rise) by anyone reproducing the study.

D Additional Results

Oracle-lite (citation identifiers without bodies). We experiment with a third information access regime. Here, the gold authorities are supplied as only identifiers rather than the full text. The results are reported in Tables 6 and 7. We observe that providing only the reference of gold authorities recovers a lot of the performance gap between closed book and

Smaller open-weights models. We additionally evaluated two smaller open-weights models: Gemma 4-26b and Qwen 3.6 - under the same three evidence-access regimes, self hosted on a Nvidia A100 GPU, using vllm at 80k context length, results in Table 6. The accuracy-SSA divergence pattern persists at smaller scale, with proportionally lower SSA ceilings: both models reach Acc@8 of 88–89% under complete oracle but pass SSA_8 at $\tau = 0.8$ on only roughly a third to two-fifths of

Model	Setting	Acc@8 (%)	SSA_8 (%)
Gemma 4-26b	Closed-book	$23.0_{[20.7,\,25.4]}$	$5.0_{[3.9,\,6.4]}$
	LiteOracle	$73.0_{[70.5,\,75.3]}$	$24.8_{[22.3,\,27.4]}$
	Oracle	$88.0_{[86.2,\,89.6]}$	$32.9_{[30.1,\,35.8]}$
Qwen 3.6	Closed-book	$1.3_{[0.8,\,2.1]}$	$5.0_{[3.9,\,6.4]}$
	LiteOracle	$72.0_{[69.5,\,74.4]}$	$32.0_{[29.2,\,35.0]}$
	Oracle	$89.0_{[87.2,\,90.6]}$	$39.1_{[36.2,\,42.1]}$

Table 6: Smaller open-weights models on HS-Bench. Subscripts are 95% CIs (Wilson for Acc@8, bootstrap for PPI++-debiased SSA_8 at $\tau = 0.8$). The accuracy-SSA divergence pattern persists, with proportionally larger gaps relative to frontier models at comparable accuracy.

products. The cleaner comparison to the frontier is the conditional pass rate $\mathrm{SSA}_8/\mathrm{Acc}_8$: at complete-oracle accuracy comparable to DeepSeek and Kimi ($\sim$84–89%), Gemma sits at 37% and Qwen at 44% conditional, against 62–64% for the two frontier open-weights models at similar accuracy. Attribution capability appears to scale with model size somewhat independently of classification capability.

Full digit-level accuracy. The body of the paper reports accuracy at the 8-digit tariff line (Acc@8), the level at which duty rates are actually assessed in U.S. customs practice. For completeness, Table 7 gives Acc@k for $k \in \{2, 4, 6, 8, 10\}$ across the four frontier models in each of the three evidence-access regimes. Two qualifications on reading the wider ladder. First, the complete-oracle regime reveals the gold 6-digit subheading together with its full 8- and 10-digit subtree by construction, so Acc@2, Acc@4, and Acc@6 are trivially 100% under that regime.

Per-authority-type stratified analysis. Table 8 breaks SSA_8 at τ=0.8 down by authority type. The pattern is the same across all four models. Products where the bench cites only the tariff schedule are easy in closed-book and barely move under oracle — the description already pins the heading. Products that turn on prior court decisions are hard everywhere; even at PureOracle, Opus sits at 58.2% and DeepSeek and Kimi hover near 50%. Oracle helps most on the slices with competing tariff lines, statutory notes, or explanatory notes to weigh (+12–18 pp for top models), which is what each authority is really doing: giving the model room to disambiguate. One caveat on EN: the WCO does not freely redistribute the Explanatory Notes, so we expose identifiers and bench-quoted paragraphs.

Model	Setting	Chapter (2d)	Heading (4d)	Sub. (6d)	Tariff (8d)	SSA_8 @ τ=0.8 (%)
*Opus 4.7	Closed-book	$82.3_{[80.2,\,84.2]}$	$71.2_{[68.7,\,73.5]}$	$60.4_{[57.8,\,63.0]}$	$45.8_{[43.2,\,48.5]}$	$\mathbf{30.5}_{[27.0,\,34.4]}$
	LiteOracle	$94.1_{[92.7,\,95.2]}$	$90.3_{[88.6,\,91.8]}$	$85.5_{[83.5,\,87.3]}$	$80.4_{[78.1,\,82.4]}$	$\mathbf{57.8}_{[51.8,\,64.9]}$
	Oracle	$100.0_{[99.7,\,100.0]}$	$100.0_{[99.7,\,100.0]}$	$99.5_{[98.9,\,99.7]}$	$93.5_{[92.1,\,94.7]}$	$\mathbf{65.7}_{[60.0,\,72.9]}$
*GPT-5.4	Closed-book	$83.7_{[81.7,\,85.6]}$	$72.2_{[69.8,\,74.5]}$	$60.9_{[58.3,\,63.5]}$	$45.7_{[43.1,\,48.3]}$	$\mathbf{24.1}_{[20.2,\,29.7]}$
	LiteOracle	$94.1_{[92.7,\,95.2]}$	$91.0_{[89.4,\,92.5]}$	$88.3_{[86.5,\,90.0]}$	$85.7_{[83.7,\,87.5]}$	$\mathbf{55.5}_{[48.2,\,63.1]}$
	Oracle	$100.0_{[99.7,\,100.0]}$	$100.0_{[99.7,\,100.0]}$	$99.6_{[99.1,\,99.8]}$	$93.7_{[92.2,\,94.9]}$	$\mathbf{65.5}_{[57.6,\,73.0]}$
*DeepSeek V4 Pro	Closed-book	$82.3_{[80.2,\,84.2]}$	$72.0_{[69.6,\,74.3]}$	$61.4_{[58.9,\,64.0]}$	$47.7_{[45.1,\,50.4]}$	$\mathbf{21.7}_{[18.9,\,26.4]}$
	LiteOracle	$92.8_{[91.3,\,94.1]}$	$89.4_{[87.6,\,90.9]}$	$87.0_{[85.1,\,88.8]}$	$84.5_{[82.4,\,86.3]}$	$\mathbf{50.0}_{[43.0,\,59.8]}$
	Oracle	$93.9_{[92.5,\,95.1]}$	$90.5_{[88.8,\,92.0]}$	$87.2_{[85.3,\,88.9]}$	$84.4_{[82.3,\,86.2]}$	$\mathbf{52.6}_{[45.8,\,60.4]}$
*Kimi K2.6	Closed-book	$74.4_{[72.1,\,76.7]}$	$61.2_{[58.6,\,63.7]}$	$44.2_{[41.6,\,46.8]}$	$29.4_{[27.1,\,31.9]}$	$\mathbf{12.0}_{[10.1,\,16.0]}$
	LiteOracle	$92.0_{[90.4,\,93.4]}$	$87.9_{[86.0,\,89.5]}$	$83.7_{[81.7,\,85.6]}$	$80.7_{[78.5,\,82.7]}$	$\mathbf{45.1}_{[38.8,\,53.4]}$
	Oracle	$94.3_{[92.9,\,95.4]}$	$90.4_{[88.7,\,91.8]}$	$86.8_{[84.9,\,88.5]}$	$82.8_{[80.7,\,84.8]}$	$\mathbf{52.7}_{[45.2,\,59.8]}$

Table 7: Per-cell accuracy at every HTSUS digit level (2-, 4-, 6-, 8-digit) and PPI++-debiased **strict support accuracy** (SSA_8) at τ=0.8 across four frontier candidate models and three evidence-access regimes. Subscripts are 95% CIs (Wilson for the direct digit-prefix accuracies; bootstrap propagating both eval-set and calibration-set variance for SSA_8). Bold marks the headline metric.

Model	Setting n =	minimal 339	competing_tariff 591	CBP 380	case_law 189	statutory_note 547	EN 532
3*Opus 4.7	ClosedBook	$54.3_{[49,60]}$	$18.8_{[16,22]}$	$21.3_{[17,26]}$	$15.9_{[11,22]}$	$19.6_{[16,23]}$	$21.6_{[18,25]}$
	LiteOracle	$72.0_{[67,77]}$	$48.3_{[44,52]}$	$48.0_{[43,53]}$	$39.7_{[33,47]}$	$52.7_{[48,57]}$	$52.3_{[48,57]}$
	PureOracle	$63.7_{[58,69]}$	$62.2_{[58,66]}$	$65.0_{[60,70]}$	$58.2_{[51,65]}$	$64.0_{[60,68]}$	$67.1_{[63,71]}$
3*GPT-5.4	ClosedBook	$39.2_{[34,45]}$	$18.8_{[16,22]}$	$18.7_{[15,23]}$	$16.4_{[12,22]}$	$17.4_{[14,21]}$	$18.4_{[15,22]}$
	LiteOracle	$68.2_{[63,73]}$	$43.9_{[40,48]}$	$46.2_{[41,51]}$	$43.4_{[37,51]}$	$50.3_{[46,55]}$	$51.0_{[47,55]}$
	PureOracle	$60.5_{[55,66]}$	$64.5_{[61,68]}$	$66.3_{[61,71]}$	$55.6_{[48,62]}$	$65.5_{[61,69]}$	$69.0_{[65,73]}$
3*DeepSeek V4 Pro	ClosedBook	$31.8_{[27,37]}$	$14.9_{[12,18]}$	$17.2_{[14,21]}$	$14.9_{[11,21]}$	$17.2_{[14,21]}$	$17.5_{[15,21]}$
	LiteOracle	$56.5_{[51,62]}$	$38.7_{[35,43]}$	$40.5_{[36,46]}$	$36.4_{[30,43]}$	$45.9_{[42,50]}$	$44.9_{[41,49]}$
	PureOracle	$51.1_{[46,57]}$	$43.3_{[39,47]}$	$46.4_{[41,51]}$	$42.9_{[36,50]}$	$54.3_{[50,59]}$	$53.6_{[49,58]}$
3*Kimi K2.6	ClosedBook	$16.7_{[13,21]}$	$8.4_{[6,11]}$	$12.1_{[9,16]}$	$9.5_{[6,15]}$	$9.2_{[7,12]}$	$11.2_{[9,14]}$
	LiteOracle	$46.3_{[41,52]}$	$34.6_{[31,39]}$	$38.5_{[34,43]}$	$35.4_{[29,42]}$	$42.4_{[38,47]}$	$44.9_{[41,49]}$
	PureOracle	$45.3_{[40,51]}$	$47.0_{[43,51]}$	$50.4_{[45,55]}$	$48.7_{[42,56]}$	$56.7_{[52,61]}$	$53.4_{[49,58]}$

Table 8: Strict Support Accuracy SSA_8 at $\tau = 0.8$ (Rogan–Gladen debiased) broken out by the six bench-authority-type slices. A product appears in slice S iff the deciding bench cited at least one authority of type S; slices overlap. Subscripts give 95% Wilson CIs computed over the subset of products in each slice. Subset sizes n are the number of products in each slice across the full 1,329-product eval set.