Tariff Classification with LLMs: What Helps, What Doesn't, and What the Standard Metric Misses

Rohan Movva
Stanford University

Abstract

Duty liability in the U.S. Harmonized Tariff Schedule is set at the 8-digit subheading; the final two digits are a statistical suffix. We verify this against the published schedule: dispersion across 3,017 8-digit groups (25,651 sibling pairs) is exactly zero, no inheriting 10-digit line differs from its parent in any duty column (0 of 11,836), and Chapter 99 trade-remedy provisions cite 8-digit subheadings 1,113 times and 10-digit codes never. The field's headline metric, exact 10-digit accuracy, therefore prices two digits with no first-order revenue consequence: on our best classifier, **53.6**% of the predictions it counts as failures cost the importer nothing, up from 34.8% on a deterministic baseline.

We then run the ablation that metric is meant to serve, on CROSS-2026: 698 leakage-audited 2026 CBP rulings, each carrying its duty rate. Retrieved evidence is what moves the system. On 100 paired items with the model at its unimpaired batch size, 8-digit accuracy is 62.0% closed book and **73.0**% with 15 retrieved candidates rendered in full beside the verbatim section and chapter notes (+11.0 pp, McNemar exact $p = 0.027$). An enforced GRI decision procedure on top of that is worth −1.1 pp on all 181 items ($p = 0.83$), an interval excluding any benefit above about four points. Across the five configurations we measured it in, on two models and two vendors, its point estimate is positive exactly once, by +2.8 pp ($p = 0.44$), in the arm where the evidence was least legible, and at or below zero in the other four, one of them significantly (−14.9 pp against the same evidence rendered in full). Nor is the retrieval effect model-independent: the retriever upgrade that helps `claude-opus-5` *harms* `gpt-5.5` by −13.7 pp, replicated over k=4 seeds at 6.8 run-to-run standard deviations from zero.

What governs whether evidence helps is not its quality but what the prompt says to do with it. We fix retrieval quality by construction, guaranteeing the gold line present or absent, and cross it with framing. The interaction is +63.1 pp on `gpt-5.5` and +72.0 pp on `claude-opus-5`, and both survive family-wise correction over all 56 tests we report. Told to choose from a list that cannot contain the right answer, both models comply on 100% of items and score 0.0%; told the same list is fallible, they leave it on most items and recover to near their closed-book accuracy. One sentence of framing, on evidence that cannot contain the answer, is worth more than any other intervention measured anywhere in this paper, and deference is a property of the instruction rather than of the model.

Three of our own results did not survive the controls we ran last, and we report the corrected figures rather than the ones that made the better story. Rendering candidates truncated at 110 characters, a choice of ours, cost 17.7 pp, more than any intervention we added. Packing 13 items into a prompt rather than 5 cost `claude-opus-5` 36.0 pp and `gpt-5.5` 0.6 pp, collapsing a 37.5 pp "capability" gap to **2.1** pp: the axis we had called competence was largely a batching axis. And a retriever upgrade worth +44.2 pp under those impairments is worth +6.3 pp ($p = 0.26$) without them. Every accuracy level in our published ladder describes a pipeline we had degraded in three ways.

A difficulty-matched contamination control finds no memorization premium (−1.1 pp, $p = 0.90$). We define *non-vacuous verification*, a predicate some wrong answer the system can plausibly produce would falsify, and apply it to seven published systems read in full text: none is vacuous outright, and the audit forced four repairs on our own definition. We release CROSS-2026 and all code.

1 Introduction

Every commercial shipment entering the United States must be assigned a 10-digit code from the Harmonized Tariff Schedule. The code determines the duty owed, whether a trade remedy applies, and whether the goods are admissible. Misclassification is not benign: under 19 U.S.C. §1592 an importer who under-declares faces retroactive liability plus penalties, while over-declaration is a silent and usually unrecovered overpayment. Classification is a legal determination made under the General Rules of Interpretation (GRI) and the binding section and chapter notes, and it is routinely contested.

The task became an active benchmark subject only recently. ATLAS (Yuvraj and Devarakonda, 2025) released the first benchmark derived from CBP's Customs Rulings Online Search System (CROSS) in September 2025; HSCodeComp (Yang et al., 2026) followed with an expert-annotated e-commerce set; Zhang et al. (2026b) published a deterministic agentic workflow for Chinese HS classification in May 2026; Xia et al. (2026) constrain decoding to legally valid paths in an HS knowledge graph; and commercial systems have been benchmarked separately (Judy, 2024). Across this literature the reported quantity is overwhelmingly exact-match accuracy at some digit depth, with 10-digit accuracy the headline where the U.S. schedule is the target.

This paper asks what that number measures. Our answer is that it substantially measures a distinction the tariff declines to make. The HTSUS is a prefix hierarchy: chapter (2 digits), heading (4), international subheading (6), U.S. legal subheading (8), statistical suffix (10). Duty is set at the 8-digit line. We verify that exhaustively against the published schedule rather than asserting it from the legal literature, and then ask what changes when errors are weighted by the consequence the schedule itself assigns them.

We also run the ablation, and we state the conditions exactly. Inference runs on two paths, and they differ in what they let us control. The Claude arms have no API credential behind them: each trial is an isolated sub-agent session receiving one batch prompt and returning JSON, audited for tool use, which is real inference from real models but leaves temperature and seed uncontrollable, so each of those conditions is a *single sample* of unknown variance. The cross-vendor arm runs over the HTTP API with no tools bound, which accepts a seed and let us repeat two conditions four times (§10.5). Everything we claim from a model is therefore a paired difference on identical items with a significance test attached, and the one claim we make about run-to-run variance rests on repeated runs rather than on a single sample. Nothing here is a claim about any system's absolute standing against published numbers. §8 states the engine and its bounds in full, and we ask that the results be read against it. This is not a paper showing that a system works. It is a measurement paper about what a system in this domain should be scored on, which of its components carry the load, and a structural ceiling that current scoring does not report.

The ablation returned the result we did not want, and then returned a second one about us. An explicit, enforced General-Rules-of-Interpretation decision procedure is the component that both the deployed system and the closest published system are built around, and no design here detects an improvement from it over plain retrieval: not on three models, not across two vendors, and not when we rebuilt the comparison with the binding notes supplied, the candidates shown in full and the batch size set where the model is unimpaired, which is the version of the test that gives structure every advantage we know how to give it (§10.3). Its point estimate there is negative at all three depths. What improves the determination is retrieval, and retrieval was the binding constraint: under the retriever we started with, the correct duty-setting line reached the model on fewer than a third of items.

The second result is that most of the retrieval gain we first measured was repair of damage we had done ourselves, and what remains does not transfer across models. Three controls we ran only after adversarial review, on candidate rendering, prompt batching and retriever quality, each removed a large part of an effect we had reported. The effect that survives is real and significant but smaller, and the same retriever upgrade that helps one model harms another. What decides the sign is not the quality of the evidence but whether the prompt tells the model that its evidence may be wrong, which we establish causally rather than by inference (§10.4).

Contributions.

1. **A structural analysis of consequence in the HTSUS, and a duty-weighted loss** (§§4, 7): zero duty dispersion among 10-digit siblings of a common 8-digit parent, no 8-digit group with two children carrying differing explicit rates, and 1,113 Chapter 99 scope references to 8-digit subheadings against zero to 10-digit codes, scoped to the 91.2% of lines reducible to an ad valorem equivalent. Scoring by consequence, 53.6% of our best classifier's 10-digit "errors" are free. The recomputation also exposes a general property: economic metrics are undefined for invented codes, so they silently condition on structural validity and must be reported beside a validity rate. The selective-prediction machinery is classical (Chow, 1970; El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017); only the loss and the deployment question are ours.
2. **CROSS-2026** (§5): 698 CBP binding rulings from 2026, built subtractively with every drop accounted for and audited for answer leakage. It is not the first CROSS benchmark: ATLAS is, and is 27× larger. Post-cutoff collection is likewise established practice (Chen et al., 2025); CROSS-2026 differs on a date partition and a resolved duty rate per item.
3. **A paired system ladder** (§9): direct → +retrieval → +structured GRI reasoning on 181 identical items, McNemar-tested. Retrieval is significant before correction (+11.0 pp, $p = 0.0045$); structure is not detectable (+2.8 pp, $p = 0.44$). §10.3 rebuilds the rungs with our own impairments removed, where they read 62.0%, 69.0% and 73.0% rather than 21.5%, 32.6% and 35.4%, and puts the structure effect at −1.1 pp on all 181 items with a 95% interval of $(-6.2, +4.0)$: the contrasts survive, the levels do not.
4. **Evidence legibility beats reasoning structure, head to head** (§9.3): re-running the retrieval rung with the same candidates shown untruncated, and nothing else changed, is worth +17.7 pp ($p = 9 \times 10^{-7}$), larger than adding retrieval at all, and beats the GRI procedure directly by +14.9 pp. It also corrects our own selection estimate from 67.9% to 89.3%.
5. **An evidence-by-model grid in which the retrieval effect changes sign** (§10): three models from two vendors crossed with four evidence settings on identical items, with `text-embedding-3-large` lifting gold 8-digit presence at k=15 from 30.9% to 48.6% and recall@50 from BM25's 31.9% to 64.3%. Upgrading the retriever is worth +33.1 pp to one model ($p < 10^{-6}$) and −13.7 pp to another (mean over k=4 seeds, 6.8 run-to-run SDs from zero). The positive half is largely an artifact of our own batching and shrinks to +6.3 pp ($p = 0.26$) once removed (§10.2); the harm replicates and the sign difference does not. Splitting each rung on whether its own retriever found the answer isolates the mechanism: on missed items the same evidence is worth +28.0 pp to the weaker model and −24.8 pp to the stronger. Recall@k does not summarize a retriever, and retrieval augmentation

has no model-independent value.

6. **Deference is causal, and one sentence controls it** (§10.4): we fix retrieval quality by construction, guaranteeing the gold 8-digit line present or absent, and cross it with whether the prompt presents the candidate list as the answer space or as fallible suggestions. The interaction is $+63.1$ pp on one model and $+72.0$ pp on another, and both survive family-wise correction. Told to choose from a list that cannot contain the answer, models comply and score 0.0%; told the list is unreliable, they leave it and recover to near their closed-book accuracy. This corrects our own earlier reading: deference is a property of the instruction, not of the model.
7. **Three controls that cost us our own headline numbers, and a measured noise floor** (§§9.3, 10.1, 10.2, E): a candidate rendering of ours worth 17.7 pp; a prompt batch size worth 36.0 pp to one model and 0.6 pp to another, collapsing a 37.5 pp apparent capability gap to 2.1 pp; and a retriever upgrade whose measured value falls from $+44.2$ pp to $+6.3$ pp once both are repaired. Re-running one condition on byte-identical prompts changes 42.3% of individual 10-digit answers while moving the aggregate 1.2 pp, which is the yardstick for which of our own deltas we may call effects. We report the corrected figures and leave the originals visible.
8. **A verification predicate that can fail, and works** (§§12, 13): only 47.5% of closed-book 10-digit codes exist in the HTSUS, and gating on that free check buys $+10.2$ pp of selective accuracy at 75.1% coverage, where the retrieval margin bought nothing; the gate's value is bounded by the invalid-code rate and shrinks to $+0.5$ pp on our best configuration, itself a finding. It is formalized by a definition: a verification predicate is non-vacuous only if a wrong-but-retrieved answer can falsify it. We apply that definition to seven published systems read in full text. None is vacuous outright, and the exercise forced four repairs on the definition itself.
9. **A bounded contamination null** (§11): 181 difficulty-matched pre-/post-cutoff pairs, closed book, no detectable memorization premium at any depth, with an interval tight enough to exclude a large effect and too loose to exclude a small one.

2 Related Work

Tariff and HS classification. ATLAS (Yuvraj and Devarakonda, 2025) is the first CROSS-derived HTS benchmark: 18,731 rulings, 2,992 unique codes, with a fine-tuned LLaMA-3.3-70B reaching 40% 10-digit and 57.5% 6-digit accuracy ahead of the frontier models it compares against. Its split is random over a corpus spanning decades, with no date stratification and no contamination analysis. HSCodeComp (Yang et al., 2026) is expert-annotated (632 e-commerce products, 27 chapters, dual annotation with senior adjudication, 2% disagreement on a re-annotated 10% sample); its best agent reaches 46.8% against a 95.0% human ceiling, and it reports the negative results that matter most here: human-written hierarchical rules *decreased* accuracy, and majority voting and self-reflection both failed. It avoids CROSS on the grounds that rulings-derived datasets leak, an accusation against ATLAS-style construction that has not, to our knowledge, been adjudicated.

Zhang et al. (2026b) give the closest system to the pipeline we audit: a six-stage deterministic workflow for Chinese HS classification combining hybrid BM25-plus-dense retrieval fused by reciprocal rank fusion, the General Interpretative Rules as explicit priority statements, exclusion-clause demotion from chapter and section notes, verbatim rule citation and per-candidate confidence, reporting 75.0% top-1 and 91.5% top-3 at four digits on HSCodeComp. They also audit 226 six-digit disagreements and report that a non-trivial fraction of HSCodeComp gold labels deviate from HS general rules, a caution we inherit since our own labels are unaudited.

We put one component of that design under a paired test, and we are careful about what the test says. Our rung C (§9) adds an enforced GRI decision procedure to a retrieval-augmented baseline; §10.3 measures it with the notes supplied, the candidates in full and the batch size unimpaired, and finds -1.1 pp on 181 items with a 95% interval of $(-6.2, +4.0)$. We did not run their system: no reciprocal rank fusion, no exclusion-clause demotion, no per-candidate confidence. What we tested is a structured-output ablation of the idea their contribution is built on, that making the GRI explicit and procedural improves the determination. The direction matches HSCodeComp's own negative results, so lightweight reasoning-side interventions on this task have now failed in two independent setups. We read that as evidence about that intervention class at currently achievable recall, not as a refutation of any published system; the much stronger retrieval stack Zhang et al. (2026b) build is exactly the part our results say carries the load.

Xia et al. (2026) enforce structural validity at decode time via a knowledge graph. §C bears on that family directly: at the recall we measure, rejecting every answer outside the retrieved candidate set would remove 47.5% of the correct answers our retrieval rung produces. Constrained decoding is safe when recall is high; we did not find a regime in which it is. Lee et al. (2024) deployed a 6-digit decision-support system with the Korea Customs Service, reaching 93.9% top-3 on 5,000 cases over 925 challenging subheadings with a 32-expert user study, following earlier work that retrieves supporting sentences as the explanation artifact (Lee et al., 2021). None of these reports duty consequence, abstention or a risk-coverage curve: ATLAS reports no confidence scoring, HSCodeComp describes no abstention, and Zhang et al. (2026b) emit confidence but develop no calibrated threshold while noting that their top-1/top-3 gap indicates errors concentrated in borderline cases. That gap is what this paper is about.

Proof-carrying and evidence-package framings. A crowded line attaches machine-checkable certificates to agent outputs, descending from proof-carrying code (Necula, 1997): action certificates in a route/review/prove loop with human approval gates (Wang, 2026); Lean 4 kernel-checked certificates over deterministic scaffolding (Koomullil,

2026); extensions to dynamic LLM outputs with a tax-computation case study (Avni et al., 2026); verifier gates before merge (Tagliabue and Greco, 2025); verified-or-abstain contracts in which a Lean kernel is the sole minter of "Verified" claims (Ren, 2026); and multi-agent legal reasoning over formalized knowledge with citation grounding and abstention (Sadowski and Chudziak, 2025). We deliberately do not adopt this framing. The term is occupied, the mechanism is published, and the decisive reason is that the deployed system we audited does not carry proofs; it carries a copy check (§12). These are architectural proposals with pilot demonstrations, leaving open whether requiring an evidence package changes correctness on an expert-labeled task. We do not answer that either.

Verification, critics, faithfulness. GLEAN (Zhang et al., 2026a) compiles expert guidelines into calibrated correctness signals with uncertainty-triggered active verification, and citation-grounded generation and its evaluation were established by ALCE (Gao et al., 2023). Against this, Huang et al. (2024) show intrinsic self-correction degrades reasoning, Cemri et al. (2025) taxonomize multi-agent failure, and Turpin et al. (2023) show chain-of-thought need not reflect the computation that produced the answer. Wang et al. (2026) name *scope laundering*, reporting conclusions consistent with a formal apparatus without having executed it, in legal reasoning specifically, and Yao et al. (2024) argue that a determination right once in k runs is not a determination, which bites where no sampling temperature is set. These are why we treat any unmeasured verification claim, including one made about the system we audited, as unsupported.

Selective prediction, contamination, legal NLP. The reject option dates to Chow (1970), with the modern theory in El-Yaniv and Wiener (2010) and deep instantiations in Geifman and El-Yaniv (2017, 2019). Learning-to-defer (Madras et al., 2018; Mozannar and Sontag, 2020) makes the human part of the decision rule. For LLMs, self-assessed confidence (Kadavath et al., 2022) (overconfident by 16–28 points here while still discriminating well above chance), abstention behavior (Wen et al., 2024), conformal wrappers (Kumar et al., 2023) and cost-sensitive conformal prediction with human-in-the-loop abstention (Singh et al., 2026) are all active. We claim no methodological novelty in any of it; what customs adds is that the cost matrix is published. Post-cutoff collection is likewise established (Chen et al., 2025; Xu et al., 2024) and is not a proof (Zhang et al., 2025); §11 turns the posture into a measurement and finds no premium to correct for. Our task sits alongside LegalBench (Guha et al., 2023), CaseHOLD (Zheng et al., 2021), LexGLUE (Chalkidis et al., 2022), statutory reasoning in tax law (Holzenberger et al., 2020; Blair-Stanek et al., 2023) and legal RAG benchmarks (Pipitone and Alami, 2024), with Dahl et al. (2024) and Magesh et al. (2025) documenting how badly legal claims fail to be grounded even in tools built to ground them. Tariff classification differs in the respect that motivates this paper: the consequence of an error is a published number.

3 Background: the HTSUS and the GRI

Hierarchy and rate forms. The first two digits of an HTSUS code give the *chapter* (e.g. ch. 85, electrical machinery), four the *heading*, and six the *subheading*, all internationally harmonized under the WCO Harmonized System. The United States adds two digits to form the *8-digit subheading* and two more to form the *10-digit statistical reporting number*; the revision we analyze contains 19,949 10-digit lines. The legally operative fact is that duty is set at the 8-digit line, the 10-digit suffix being a statistical reporting distinction maintained for trade-data purposes. Our contribution in §4 is not to state this, which is customs-law background, but to verify it exhaustively and to check whether anything else observable in the schedule attaches below 8 digits. Of the 19,949 lines, 7,423 are Free and 10,337 carry an ad valorem percentage, while 1,380 carry specific rates (cents per unit), 429 compound rates and 380 another or unparsed form. Only rates reducible to an ad valorem equivalent are commensurable across lines without unit prices we do not have, leaving 18,189 lines (91.2%); the excluded 8.8% is concentrated in agriculture and textiles, where classification disputes are common. That is the most important scope limit on everything below.

The GRI. Classification is governed by six General Rules of Interpretation applied in order. GRI 1 is operative in most cases: classification is determined by the terms of the headings and by any relative section or chapter notes, which are binding text and frequently exclude goods the heading's ordinary language would cover. GRI 3 applies when goods are prima facie classifiable under two or more headings: 3(a) prefers the most specific description, 3(b) classifies mixtures, composite goods, and retail sets by the component giving them their *essential character*, and 3(c) is a residual tie-break; GRI 6 applies the same logic at subheading level. Two properties matter. First, GRI 1 and GRI 6 are largely lookup-and-read operations over text, whereas 3(a) and especially 3(b) are defeasible gradient judgments with no mechanical decision procedure; "essential character" has no formal adjudicator. Second, the notes are the instrument that resolves the hardest boundaries: whether an article is electrical (ch. 85) or mechanical (ch. 84) machinery is a notes question, not a vocabulary question.

4 The Metric Problem

4.1 Duty-rate dispersion within the hierarchy

We take the published Column 1 General rate for every 10-digit line, reduce it to an ad valorem equivalent where possible, and measure dispersion among sibling lines sharing a common ancestor at each depth. No model is involved; this is a property of the schedule (Table 1, Table 1). The 3,017 8-digit groups counted there are those whose rates all reduce to an ad valorem equivalent, out of 3,274 8-digit groups with two or more 10-digit children in the raw schedule.

Sibling 10-digit lines sharing a common…	Groups	Pairs	Identical (% groups)	Mean	p90	Max
				|Δ duty rate| (pp)		
8-digit subheading	**3,017**	**25,651**	**100.0**	**0.00**	**0.0**	**0.0**
6-digit subheading	3,081	79,479	54.6	4.99	12.1	350.0
4-digit heading	1,065	449,095	36.8	4.65	12.0	350.0
chapter	97	5,293,451	6.2	3.81	8.9	350.0

Table 1: Duty-rate dispersion among sibling 10-digit lines, over the whole schedule. Within an 8-digit subheading every group is internally identical: 3,017 of 3,017 groups, 25,651 of 25,651 pairs, maximum difference 0.00 percentage points. Dispersion appears only once the shared ancestor is 6 digits or shallower.

4.2 Ruling out rate inheritance as an artifact

The obvious objection is that this is an artifact of publication convention. USITC leaves the rate cell blank on 10-digit lines whose rate is set at the 8-digit parent; 11,836 of 19,949 lines inherit this way and 8,112 carry a rate in their own row. If sibling agreement were merely inheritance we would expect 8-digit groups in which two or more children carry their own *explicit* and differing rates. There are none: 0 of all 3,274 multi-child groups, anywhere in the schedule, in any rate form.

We then asked whether any other consequence *observable inside the HTSUS* attaches below 8 digits. The Column 1 General, Column 2 and Special (FTA preference) columns are identical to the 8-digit parent for all 11,836 inheriting lines, with zero exceptions in all three. Chapter 99, which carries the Section 301 and Section 232 measures, delimits scope by 8-digit subheading 1,113 times across 267 unique subheadings and by 10-digit code exactly zero times. Restricted to the 698 benchmark true codes the same gradient holds: 100% zero spread within 8-digit groups ($n = 233$ codes with siblings), 50.0% within 6-digit (mean 4.24, max 45.0 percentage points), 31.0% within 4-digit (mean 7.40, max 131.8) and 2.3% within chapter (mean 16.07, max 163.8).

4.3 What this does and does not license

The claim is that tariff treatment, as expressed in the HTSUS, is constant below the 8-digit subheading, for the 91.2% of lines whose rate reduces to an ad valorem equivalent. The broader claim would be wrong. We did not test, and assert nothing about, any of the following: **(i)** antidumping and countervailing duty scope orders, administered by Commerce outside the HTSUS and defined by written product description, with the HTS numbers they list explicitly non-dispositive; **(ii)** partner-government-agency admissibility, where FDA, USDA, FCC and other flags live in the ACE/ITDS message set and some flagging is known to operate at the 10-digit line, so a 10-digit error can cause an admissibility failure with no duty consequence at all; **(iii)** absolute and tariff-rate quota administration; **(iv)** non-ad-valorem rate forms (8.8% of lines); and **(v)** realized revenue. We measure rate *dispersion*, not dollars, and a 350-point dispersion in a group under which nothing is imported is economically irrelevant. A customs broker reading only our headline would object within thirty seconds, correctly. The finding is worth considerably more scoped than unscoped.

4.4 Consequence for the field's headline metric

Exact 10-digit accuracy is the dominant reported number: ATLAS headlines 40% fully-correct 10-digit (Yuvraj and Devarakonda, 2025) and HSCodeComp reports 10-digit accuracy as its primary measure (Yang et al., 2026). By the structure of the schedule, the last two digits of that string carry zero first-order revenue consequence. A system evaluated at 10 digits is therefore scored substantially on a distinction that costs the importer nothing, while the economically decisive 8-digit boundary is reported as a secondary figure or not at all. This is not a claim that the 10-digit code is unimportant; it is legally required on the entry summary, and admissibility and quota consequences may attach to it. It is a claim about what an accuracy number means: two systems with equal 8-digit and unequal 10-digit accuracy differ in statistical reporting quality, not in duty exposure, and a metric that cannot separate those reports a blend of them.

5 CROSS-2026

Why another benchmark. We claim neither the first CROSS benchmark (ATLAS does that, is 27× larger and preceded us by eleven months) nor post-cutoff collection as a method (Chen et al., 2025). CROSS-2026 exists because our two measurements need what ATLAS and HSCodeComp do not supply: a date partition (so a contamination-premium experiment is possible *on CROSS*, the corpus HSCodeComp accuses of leakage) and a resolved ad valorem rate per item (so a duty-weighted loss can be computed). Table 2 states the comparison, including where we are worse.

Construction. We harvested 1,479 rulings issued between 2026-02-11 and 2026-08-07 from CBP's CROSS endpoint and built the benchmark subtractively; Table 9 (appendix) accounts for every dropped item. The result has 698 items, 389 distinct 10-digit labels, 346 distinct 8-digit labels and 59 chapters, with median description length 705 characters (mean 820; p10 285, p90 1,443) and a chapter distribution dominated by ch. 85 (81 items), ch. 95 (69), ch. 84 (67) and ch. 39 (49). Labels are CBP's own binding determination as published; no human re-annotation was performed, which we regard as a real weakness rather than a design choice. The label distribution is markedly non-uniform: the 20 most frequent labels cover 193 items (27.7%), and one alone (festive articles, 9505.90.6000) covers 37 near-identical rulings. The consequences for inference are stated in L1.

Leakage control: a failure, its replacement, and what still survives. A CROSS ruling carries the answer and CBP's reasoning in the same document as the product description, and removing them is harder than it looks. Our first attempt, token-level scrubbing of code patterns and classification vocabulary, failed inspection in two ways: partial suffixes such as `.4590` do not match a full-code regex but are recognizable statistical suffixes, and, more damagingly, CBP's analytical prose survived, including "we disagree with the suggested classification" and "considered a set per GRI 3(b)," which leak the reasoning and in the second case the operative rule. We discarded token scrubbing and now truncate each ruling at the first sentence containing any classification signal, keeping only preceding text. This is lossier: 316 of 698 kept items lost at least one sentence (mean 3.54, median 1, max 37; 382 needed none), and it is why 145 rulings were dropped for having no usable description left. An audit of the finished set finds 0 items containing an HTS-code pattern and 0 containing their own answer's 4-digit heading; three flagged items were manually inspected and all three were false positives from part numbers and SKUs (`117-00063.000.A1`, `VX 8806.000`, `305969 26-26X004`).

The audit's third clause was too strong, and we restate it. An earlier version of that sentence claimed zero items retain classification vocabulary. Scanning instead for CBP's determinative *voice* ("this office", "we disagree/agree/find/note/believe", "you suggest", "essential character", "composite good", "put up for retail sale", "principal function/use") returns 40 of 698 items (5.7%), e.g. "the decorative wire pumpkin is composed of different components ... and is considered a composite good," which is GRI 3(b) reasoning rather than description. Eight of those items fall in the 181-item ladder subset, and rung C is correct at the 8-digit line on 5 of the 8 against 35.4% overall. The cell is too small to interpret, but it points the direction a leakage concern predicts. The truncation removes the answer and the heading; it does not remove every trace of the adjudicator's reasoning, and 5.7% is the honest number.

Contamination posture, and how far it goes. All items post-date 2026-02-11, which is a control a random split over a decades-spanning corpus does not provide. It proves less than it appears to, in three ways that we quantify where we can. The posture is model-relative and must be reported per model: against the model evaluated here, whose stated cutoff is May 2026, only 259 of 698 items (37.1%) are strictly post-cutoff, so the benchmark as a whole is a date-partitioned set rather than a post-cutoff one. CBP issues many rulings on near-identical goods, so a post-cutoff ruling may have a pre-cutoff twin (Zhang et al., 2025). The schedule itself is public and heavily reproduced, so temporal holdout on *rulings* does not create temporal holdout on *the answer space*: a code first published in 2022 can be the answer to a 2026 ruling. We do run the experiment that would establish a contamination premium, on difficulty-matched pairs drawn against this partition, and it finds none (§11); that weakens the motivation for the date partition rather than vindicating it, and we say so there.

Selection bias. CROSS rulings exist because an importer asked, and importers ask when the classification is unclear or the stakes are high. CROSS-2026 is therefore conditioned on ambiguity and is not representative of the entry-line distribution a deployed system sees, most of which is routine repeat cargo. That makes the benchmark usefully hard, and makes any duty-weighted statistic on it a statistic about *contested* goods. Those are plausibly goods with larger duty spreads, which would inflate our duty-error figures relative to a real entry stream. We have the data to test this and did not. Separately, all 698 items are from the NY letter track: 91 harvested HQ-track rulings, the harder and more contested track, were lost in extraction (L2), so the benchmark's difficulty ceiling is set by a construction failure rather than by design.

	CROSS-2026 (this work)	**ATLAS** (Yuvraj and Devarakonda, 2025)	**HSCodeComp** (Yang et al., 2026)
Size	698 items	18,731 rulings; 200 val / 200 test sampled	632 items
Source	CBP CROSS binding rulings, 2026-02-11 to 2026-08-07	CBP CROSS, corpus spanning decades	E-commerce product listings
Label provenance	CBP's binding determination; **no human annotation performed**	CBP ruling code	Dual expert annotation, senior adjudication
Label-noise audit	**None**	None reported	2% on a re-annotated 10% sample; contested by Zhang et al. (2026b)
Split strategy	Date-partitioned (2026 only); 37.1% strictly post-cutoff for the model evaluated here	Random sample; no date stratification	Single expert set; avoids CROSS
Within-item leakage control	Sentence truncation at first classification signal; 0 codes, 0 headings, 5.7% residual evaluative language	Not described	N/A (source contains no answer)
Difficulty selection	Request-driven, skewed to ambiguous goods; NY track only, no HQ track	Same request-driven skew	Curated for expert-level difficulty
Duty consequence per item	**Yes**	No	No
Abstention / coverage	Supported and scored	Not reported	Not described

Table 2: CROSS-2026 against the two established benchmarks. ATLAS is substantially larger and prior; HSCodeComp has a real annotation protocol and a human ceiling we do not have. CROSS-2026 differs on the date partition and the duty label, and those two axes are its entire justification.

6 Retrieval as a Ceiling

The pipeline we audited (§12) retrieves the top 50 tariff-line candidates, renders them into the prompt, and validates that the emitted code is a member of that candidate set, retrying once on failure; Xia et al. (2026) take a stronger version by constraining decoding to graph-valid paths. Under any such design retrieval recall@k is a *hard upper bound* on end-to-end accuracy, since an answer never retrieved cannot be emitted. Reasoning improvements in such systems are therefore measured against a ceiling nobody currently reports.

We index all 19,949 10-digit lines with ancestry-resolved legal text and query with the benchmark descriptions using deterministic Okapi BM25 ($k_1 = 1.5$, $b = 0.75$). For 336 of 698 items (48.1%) the correct 10-digit line never appears within the top 500, and where found the median rank is 50. At $k = 50$, recall for the correct 6-digit subheading (37.4%) exceeds exact 10-digit recall (26.1%) by 11.3 points, the signature of a retriever that finds the right neighborhood more often than the right line. Reachability predicts downstream behavior: restricted to items whose gold line appeared anywhere in the top 500, the BM25 top-1 baseline's chapter accuracy is 50.3% against 17.3% where it never appeared, with mean duty error 2.76 versus 3.91 percentage points. Retrieval failure is not a separate failure mode from classification failure here; it is substantially the same one.

Dense retrieval, and a prediction of ours that it falsified. Product-description-to-legal-text matching is a textbook semantic-gap task, so a lexical retriever doing badly at it is close to a null result on its own. We pre-registered the objection as prediction F1: dense recall@50 for the correct 8-digit line would exceed BM25's 31.9% by at least 20 points, and a gain within 10 points would falsify our ceiling framing. Indexing the same 19,949 lines with `BAAI/bge-small-en-v1.5`, a 33M-parameter open bi-encoder, on CPU and querying with the same 698 descriptions gives the released retrieval artifact.

F1 is falsified. The measured gain at $k = 50$ is $+10.4$ pp (31.9% $\rightarrow$ 42.3%), exactly at the boundary we set in advance and far short of the ≥ 20 predicted. Dense retrieval helps at every k and helps more as k grows, but it does not rescue the ceiling: at the deployed top-50 operating point a strong dense retriever still fails to surface the correct 8-digit line for 57.7% of items, and the exact 10-digit line is absent from the top 500 for 250 of 698 items (35.8%). The ceiling is a property of the task, not an artifact of lexical matching.

The remaining caveat, at full strength. `bge-small-en-v1.5` is a strong open baseline, *not* the production `voyage-3-large` the audited system uses; no embedding API key was available. A better retriever would move these curves up by an unmeasured amount. We claim the direction and the order of magnitude, and placing the production retriever on this curve remains the cheapest experiment left undone (P8). Separately, the repository's own diagnostic reports that 20 of 37 investigated stubborn failures had the correct code within the top 50; that conditions on failures and is not comparable to our figure, so we do not treat it as a measurement.

7 Duty-Weighted Evaluation and Selective Prediction

The loss. For predicted code p and true code t, define $\ell_{\text{duty}}(p,t) = |\, r(p) - r(t)\,|$, where $r(\cdot)$ is the Column 1 General ad valorem rate resolved through the hierarchy (a 10-digit line inherits its parent's rate; by §4 the inheritance is exact). It is denominated in percentage points of entered value, the unit in which 19 U.S.C. §1592 liability scales. Items whose true or predicted rate is not ad-valorem-resolvable are excluded and counted, leaving $n = 636$ of 698 duty-scorable. We report signed error separately, because the directions are not symmetric: under-declaration creates penalty exposure, over-declaration is an unrecovered overpayment.

One baseline, scored both ways. We score BM25 top-1, a deliberately weak and fully deterministic reference point that is neither the audited system nor competitive with published work. Its absolute accuracy is not a finding; the point is that the two metrics characterize the *same* predictions differently on the deterministic baseline. We repeat the analysis on the ladder's own predictions in §F, where the finding strengthens; the two versions are reported side by side deliberately.

Of the 617 duty-scorable predictions wrong at 10 digits, 215 (34.8%) carry exactly zero duty consequence: roughly a third of what the headline metric counts as failure is economically free. That is not an artifact of the statistical suffix alone. Of the 595 duty-scorable predictions wrong at the 8-digit line, 193 (32.4%) also carry zero duty error, because 47.4% of the benchmark's true lines are duty-free and rate collisions between distinct subheadings are common. A large share of classification errors here are errors of *reporting*, not of *liability*. The converse matters more for deployment: error mass concentrates, with the worst decile carrying 48.1% of total duty error, 36.8% of predictions exactly right on rate and only 2.0% above 20 points. Directionally, 225 predictions (35.4% of scorable) under-declare at mean 5.18 points and 177 (27.8%) over-declare at mean 5.29; a compliance officer cares about those 225 specifically, and exact-match accuracy cannot find them.

A statistical argument for the same change of metric. A second reason to prefer the duty-weighted outcome has nothing to do with customs law. At $n = 698$ a McNemar test on a binary exact-match outcome has a minimum detectable effect of roughly 3.4–5.8 percentage points at 80% power uncorrected, over discordant-pair shares of 0.10 to 0.30 (`code/mde.py`), larger than most deltas this literature reports between adjacent configurations. The continuous duty-weighted outcome gives one roughly an order of magnitude smaller on the same sample; we quote it as an order of magnitude rather than a figure, because the per-item duty-error dispersion needed to pin it down is not among our committed artifacts. A benchmark of this size can resolve differences in duty consequence that it cannot resolve in

exact-match accuracy.

Error structure. The baseline's predictions land as follows. Of the 656 predictions wrong at the 8-digit line (all 698 as denominator), 198 (30.2%) are nonetheless in the correct chapter and 86 (13.1%) within the correct 4-digit heading: the error is frequently local, not categorical. The most frequent cross-chapter confusions are substantively hard boundaries rather than noise, and the ch. 84/ch. 85 pair runs both ways (23 items). That pair is the canonical Section XVI problem, precisely what chapter and section notes exist to resolve. Two further observations. First, Chapter 98 (U.S. goods returned, temporary imports) is a lexical attractor, predicted 25 times against a true count of 1; landing there is structurally inadmissible for an ordinary commercial import rather than a near miss, a class of error exact-match accuracy scores identically to a one-line miss. Second, longer descriptions help only at coarse depths: accuracy varies with description length at chapter (Mann–Whitney $p = 0.0001$) and 4-digit level ($p = 0.0024$) but not at 8 digits ($p = 0.60$). More text buys the right neighborhood, not the right line.

Selective prediction, and a negative result. The framework is standard: given a confidence signal $s(x)$ and threshold τ, answer when $s(x) \geq \tau$ and defer otherwise; sweeping τ traces a risk-coverage curve (Chow, 1970; El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017). The domain contribution, if any, is only that the risk axis can be denominated in percentage points of entered value rather than $0/1$ error, making the operating point a statement about money. On this deterministic baseline the only available signal is the retriever's own score margin (top-1 minus top-2 score); we ranked all 698 predictions by margin and swept coverage. Two further signals become available once a model is in the loop, and both are examined later: the model's self-reported confidence (§D) and a structural check on the emitted code (§13).

This is a negative result and we report it as one. A usable confidence signal produces a steeply decreasing risk curve as coverage shrinks; neither curve does. Accuracy rises from 6.0% at full coverage only to 10.0% at 10% coverage, moving non-monotonically in between ($7.5 \to 8.0 \to 6.3 \to 6.7$), and duty-weighted risk trends correctly from 3.30 to 2.54 pp as coverage tightens from 100% to 25% before reversing to 2.83 pp at 10%. Total movement is smaller than the bootstrap interval on the full-coverage mean. BM25 score margin is not a usable abstention signal for this task. (An earlier draft asserted that the duty-weighted curve was monotone where the accuracy curve was not; plotting it showed the claim false and we removed it rather than softening it.) The obvious confound is that a 6.0% classifier may be too weak for *any* signal to separate, and it is settled later, against our earlier reading: §13 runs selective prediction on a real model at 35.4% accuracy and finds two signals that work, a structural check on the emitted code (AUROC 0.630–0.668) and the model's own confidence (AUROC 0.626–0.683), every interval excluding chance. What fails here is the retriever's score margin specifically, not the possibility of a signal.

8 Inference Engine and Its Limitations

This section exists so that no reader can mistake what produced the model numbers in §9 and §11. It should be read before either.

What the engine is. No `ANTHROPIC_API_KEY` was available. We ran inference by dispatching isolated sub-agent sessions of the session model (`claude-opus-5`), each receiving exactly one batch prompt and returning JSON, and we scored the output offline with committed code; the cross-model replication of §9.5 used the same path with `claude-haiku-4-5`. This is real inference from real models, held constant within every comparison. It is not an API harness, and four differences bound what can be claimed.

(i) Temperature and seed are not controllable, so every condition is a *single sample* of unknown variance. Two things follow. Statistically, pairing is a defense rather than an apology: under the null that two conditions have identical per-item success probabilities, independent single draws make the discordant cells exchangeable, so McNemar's conditional exact test is valid at level α even though every cell is one sample. What pairing does not do is recover power, or control anything that differs between the conditions besides the intervention. We also measured the instability rather than arguing about it: re-running one condition on byte-identical prompts moved the aggregate 8-digit score by 1.2 points while changing 42.3% of individual 10-digit answers (§E). That is what licenses reporting $+11.0$ pp as an effect and what forces us to report $+2.8$ pp as undetectable at this resolution rather than as zero; it is one observation of the swing, not an estimate of its variance (P4b). Yao et al. (2024)'s objection, that a determination right once in k runs is not a determination, applies to our numbers exactly as we apply it to others'.

(ii) Items are batched. Prompts carried 5–24 items each (24 for the contamination arm, 13 for the ladder rungs, 5 for the notes and batch-size arms), order shuffled deterministically (seed 20260819), with the two members of each contamination pair interleaved across batches so that era is not confounded with batch position or session. Batching is a deviation from one-item-per-call inference and can induce order effects; §9.1 shows one in our own data, and L12 states what we did and did not control.

(iii) Closed book is instructed *and audited*, not assumed. We ran 290 trials on this path: 16 contamination, 14 each for rung A (the controlled re-run of §9.1), rung B, rung C, rung B′ (§9.3) and rung B$^+$ (§10), 6 repeated rung-C batches for the reliability measurement of §E, 14 for the cross-model replication of §9.5, 50 for the two arms of §9.4, 20 for the closed-book batch-size control of §10.1, 20 for the retrieval arm of the batch-5 grid (§10.2), 20 for the deference manipulation (§10.4) and 74 for the two arms of the fair ladder (§10.3). The tool-use count and the target of every call in every trial were recorded. Across all 290 trials, every call was a read of that trial's own batch file, and no trial read anything else, ran a search, or reached the network; a trial whose calls did not satisfy that predicate would have been

excluded from scoring rather than reported, and none was. Counts are not uniform: trials whose prompts fit one read window made exactly one call, while the rung-B′ trials (prompts to 84 kB) made one to three and the notes-arm trials (to 235 kB) made two to five. We report the distribution rather than the "exactly one call" invariant an earlier draft stated, because the invariant we can actually defend is the target of the calls, not their number. Nothing in the transcripts shows a lookup of the answer. The cross-vendor arm of §9.6 runs outside this path entirely: 329 HTTP calls across six rungs, the deference conditions and four repeat seeds, with no tools bound, so there is no tool-use audit to report because there was no tool surface to audit. Each logged call carries the model snapshot the API returned (`gpt-5.5-2026-04-23` on all 299 successful calls), the parameters it accepted (it rejected `temperature` and accepted a seed, identically on all of them), and its token counts. The reconciliation is not clean and we report the gap rather than the headline: 321 calls are logged, 299 of them successful and 22 failed, against 329 response files on disk. The 30 unlogged files are from two runs of the k=4 replication that were terminated before the manifest was written, at a time when it was written only on clean exit. Their response payloads are intact and are what the analysis reads; what was lost is their per-call metadata. The harness now flushes the manifest after every call, and `data/openai_run_accounting.json` carries the reconciliation.

(iv) Answer-lookup vectors were removed from the prompt surface. Items carry opaque identifiers (`i0001...`) rather than CBP ruling numbers, gold labels live outside the directory the prompt files sit in, and citations to prior rulings inside the retained descriptions are masked as `[PRIOR RULING]`. These are precautions, not proofs: we cannot demonstrate that a model with no network access and one file read did not recall an answer, only that we did not hand it one.

What follows from this. Every model result below is a within-model, within-item comparison. None is a claim about a system's absolute standing against ATLAS, HSCodeComp or the audited pipeline, and none should be quoted as one: the numbers are low, the benchmark is hard and request-conditioned, and the prompts are ours rather than anyone else's. Where we describe a deployed system (§12) the description is a static code audit, and we report that system's accuracy nowhere: in the checkout we examined, the raw evaluation artifacts behind its reported numbers are absent, and two of its own reports disagree by six cases on the same 97-item gold set under the same prompt and model.

9 The System Ladder: Retrieval Helps; Structure Is Not Detectably Better

The question an ablation should answer is which component of an agentic classifier carries the load. We built the smallest ladder that separates the two candidates most of this literature invests in (retrieval and structured legal reasoning) and ran all three rungs on the same 181 CROSS-2026 items with the same model, fully paired; §9.5 then replicates the decisive contrast on a second model.

A **direct** closed book: product description → code, no candidates.

B +**retrieval** A, plus the top 15 dense-retrieved HTSUS lines (§6) rendered into the prompt.

C +**structure** B, plus an enforced GRI decision procedure (material, principal function, the GRI actually relied on, code written last) and a structured output.

All three rungs run on the same 181 items, in the same order, at the same number of items per prompt. We did not have that property on the first attempt and had to go back and establish it (§9.1). Comparisons use McNemar's exact test on the paired outcomes with Holm's step-down correction *within* each digit depth; accuracy intervals are item-level percentile bootstraps (2,000 resamples, fixed seed), and L1 in §14 applies to them.

Rung	What it adds	correct	8-digit acc.	95% CI
A direct	closed book, no candidates	39	**21.5%**	(15.5, 27.6)
B +retrieval	top-15 dense candidates in prompt	59	**32.6%**	(25.4, 39.2)
C +structure	B + enforced GRI procedure, structured output	64	**35.4%**	(28.2, 42.5)

Comparison	Δ (pp)	95% CI	disc. b/c	p (exact)	p (Holm)	
A → B	+11.0	(+3.9, +18.2)	13 / 33	0.0045	**0.0091**	**significant**
B → C	+2.8	(−2.8, +8.4)	11 / 16	0.4421	0.4421	not significant
A → C	+13.8	(+6.2, +21.4)	14 / 39	0.0008	**0.0024**	**significant**

Table 3: The ladder at the duty-setting 8-digit line, $n = 181$ paired items, `claude-opus-5`. **Top:** accuracy per rung with item-level bootstrap intervals. **Bottom:** paired McNemar exact tests, Holm-corrected within depth; b/c are the discordant counts (first-only correct / second-only correct). Adding retrieval is significant; adding an enforced GRI procedure on top of it is not.

Retrieval carries the measurable gain. A → B is +11.0 pp ($p = 0.0045$ exact, 0.0091 after Holm, significant); B → C is +2.8 pp ($p = 0.44$), not detectable at this sample size. The cumulative A → C effect, +13.8 pp ($p_{\text{Holm}} = 0.0024$), is real, but the decomposition assigns nearly four fifths of it to the retrieval step. The enforced GRI procedure is what the deployed system's prompt is organized around and the idea the closest published work is built on (Zhang et al., 2026b); here, adding it on top of retrieval produces no gain we can distinguish from noise.

That statement is easy to overstate in either direction, so we are exact about it. A non-significant +2.8 pp with a 95% interval of $(-2.8, +8.4)$ is not evidence that structured GRI reasoning does nothing; it is a *failure to detect*. The

detection floor is real: re-running one condition on byte-identical prompts moved the aggregate score 1.2 pp (§E). The A $\rightarrow$ B effect is roughly nine times that swing, which is why we call it an effect; the B $\rightarrow$ C delta is about twice it, from a single pair of runs. The claim we defend is "no detectable benefit, and this design could not reliably detect one of this size," not "structure does not help." Reading $+2.8$ pp as an improvement would over-read our data; reading it as a refutation would too. Our rung C is also a lightweight version of the idea, not a reimplementation of any published system, and it does not put the chapter and section notes in the prompt (P13).

At other depths nothing is significant, and retrieval may even hurt (Table 3). Adding candidates moves 6-digit accuracy -3.9 pp and chapter accuracy -1.7 pp; neither is close to significance ($p = 0.32$, $p = 0.68$), and both shrank when we removed the batch-size confound, which is itself informative about how much of the original signal was context load. We flag them, and claim nothing, because their sign is consistent across two depths and has a plausible mechanism: a candidate list can anchor the model into a wrong-but-lexically-plausible neighborhood it would not have chosen unaided, trading coarse correctness for fine-grained plausibility. If real, that is a genuine cost of retrieval augmentation, and it is invisible to a benchmark reporting only the deepest depth. In the other direction, all four B $\rightarrow$ C deltas are positive ($+2.8$, $+5.0$, $+3.9$, $+4.4$), which is weak evidence for a small real structure gain that our n cannot resolve; consistency requires reading that pattern as seriously as the negative one. Both deserve an n that can resolve 4 points.

9.1 A confound in our own design, caught by review and removed

The numbers above are from the second version of this experiment. The first was confounded, we did not catch it ourselves, and the episode belongs in the paper rather than in a changelog.

What was wrong. Rung A was originally not a separate run at all: we reused the closed-book predictions from the modern arm of the contamination experiment (§11), reasoning that the items and the model were identical and a closed-book condition is a closed-book condition. It is not. Those batches carried 23.9 items per prompt against 12.9 for rungs B and C, so A $\rightarrow$ B compared retrieval against no-retrieval and short context against long context at the same time. The confound was not hypothetical: rung A shows a within-batch position effect, 26.0% correct at the 8-digit line in slots 0–12 against 16.9% in slots 13–23. Restricting the original comparison to rung A's first 13 slots, the crudest available control, shrank the effect from $+10.5$ pp to $+6.7$ pp and took it out of significance ($p = 0.281$).

What we did, and what happened. We re-ran rung A from scratch with batch composition *byte-identical* to rungs B and C, differing only in the absence of the candidate block: same items, same order, same 13 per prompt. We discarded the reused predictions; the re-run added 14 audited trials, and the confounded responses are retained in `data/responses/p3_*`. The effect survived and strengthened slightly: A $\rightarrow$ B moved from $+10.5$ pp ($p = 0.0094$) to $+11.0$ pp ($p = 0.0045$), and rung A's own accuracy from 22.1% to 21.5%. B $\rightarrow$ C is untouched, neither rung having changed. The two non-significant negative deltas at 6-digit and chapter both shrank, from -4.4 and -3.9 points to -3.9 and -1.7, which suggests a meaningful share of the original "anchoring cost" signal was context length.

We report this at length because the effect could have gone the other way, and because the confound is a generic hazard: conditions that differ in prompt length differ in more than the thing being ablated, and reusing a condition across two experiments is exactly how prompt length stops being held constant without anyone deciding that it should.

9.2 Decomposing the failure

This is the part of the result we would keep if we could keep only one. Of the 181 items, the correct 8-digit line was present in the 15 candidates the model was actually shown for only 56, or 30.9%. Splitting every rung on that single fact (Table 10 (appendix), and the right panel of Table 3) separates a retrieval term from a selection term.

Given the correct candidate, rung B selects it 67.9% of the time and rung C 71.4%. Without it they land at 16.8% and 19.2%, below the 21.5% the model reaches closed book: those answers are recovered from parametric knowledge in spite of a list that does not contain the right line, and the shown list appears to cost slightly more than it gives on that subset. Retrieval is the binding constraint; selection is comparatively strong. This is the model-level confirmation of §6: an intervention on the reasoning side optimizes the healthier of the two terms, and the arithmetic of the ladder is roughly 0.309×0.679 plus a parametric-recovery remainder. It sets the price of the missing experiment too: the highest-value change to this pipeline is better recall, and the second is measuring how far a production retriever moves 30.9%.

A confound inside the selection term. The 67.9% selection rate is measured under a candidate rendering that truncated each line's legal text at 110 characters, which removes the leaf description (the text that distinguishes a line from its siblings) from 77.4% of the 2,715 candidate lines the ladder actually showed (L5). So "retrieval recall is the ceiling" and "our candidate rendering is the ceiling" are not separated by rungs A–C, and the selection term should be read as a lower bound. §9.3 runs the separating experiment. The rendering turns out to matter more than any other variable we manipulated.

9.3 Rendering the retrieved evidence in full

Rung B′ re-runs rung B with the candidate text untruncated and everything else held fixed: the same 181 items in the same order, the same 15 candidates per item in the same order, the same description budget, the same task instructions, and the same batch composition, verified byte-identical across all 14 key files. The only difference is how much of each retrieved line the model was permitted to read, a mean of 105.3 characters per candidate under rung B against 296.5 under B′, because 85.3% of the 2,715 shown lines are longer than the 110-character window rung B applied.

	Accuracy (%)					
Depth	B (110 char)	B′ (full)	Δ	95% CI	McNemar p	b/c
6-digit	54.1	66.3	+12.2	(6.1, 18.2)	2×10^{-4}	6/28
8-digit (duty-setting)	**32.6**	**50.3**	+17.7	(11.0, 24.4)	9×10^{-7}	**6/38**
10-digit (full code)	16.6	41.4	+24.9	(18.2, 31.5)	4×10^{-12}	2/47

Table 4: Rung B′: the same retrieval, shown in full. Paired on the same 181 items, exact McNemar, b/c the discordant cells. The +17.7 pp gain at the duty-setting line is larger than the +11.0 pp that adding retrieval at all bought over the closed-book rung, and it required no new retrieval, no new reasoning machinery, and no new model call beyond the re-run itself.

The effect is the largest this paper measures on the primary model (Table 4 and Table 4). At the duty-setting 8-digit line accuracy rises from 32.6% to 50.3%, a paired gain of +17.7 pp (95% CI (11.0, 24.4), exact McNemar $p = 9 \times 10^{-7}$, discordant cells 6/38); at the full 10-digit code it rises from 16.6% to 41.4%.

What this does to the selection estimate. Conditioned on the gold 8-digit line being among the 15 shown, selection rises from 67.9% to 89.3%. The ≈ 70% figure of §9.2 was therefore substantially an artifact of our own prompt rather than a property of the model, and we restate it as such. Given evidence that is both retrieved and legible, selection is close to solved on this sample. Presence is not, and is unchanged at 30.9% by construction, because B′ retrieves exactly what B retrieved.

Legible evidence beats explicit procedure, head to head. Rungs C and B′ are the two interventions available on top of one retrieval result: C adds an enforced GRI decision procedure to the truncated evidence, B′ shows the same evidence in full and adds no reasoning machinery at all. On the same 181 items, B′ beats C by +14.9 pp (35.4% against 50.3%; 95% CI (7.8, 22.0), exact McNemar $p = 9.9 \times 10^{-5}$, discordant cells 10/37). This is the sharpest available form of the paper's claim, and it is a direct comparison rather than an inference across rungs: on this task, at this recall, making the retrieved evidence legible is worth substantially more than making the reasoning procedure explicit.

Three things this does not show. It does not lift the recall ceiling. The gold line is still absent for 69.1% of items, and B′ reaches 89.3% where it is present against 32.8% where it is not, so the binding constraint identified in §9.2 is unchanged and is now more starkly located. It does not convert the B→C null into a positive result about structure, and it does not test structure fairly either: rung C has still never been run against untruncated evidence, so the right reading is that the structure intervention was evaluated on evidence we had crippled, and P13 asks the matching question for chapter and section notes. Finally, structural validity is essentially unmoved (75.7% against 75.1%), so the gain is in choosing better among real codes rather than in inventing fewer of them.

What it says about the rest of this paper. This is the second confound in our own design that came out of adversarial reading rather than out of our own checking: the first was batch composition (§9.1), this is candidate rendering. In both cases an earlier draft named the confound, declined to resolve it, and reported a number anyway; in both cases running the control moved the number materially. We record the pattern because it is the best available evidence about how much weight the uncontrolled items still listed in §14 should carry, and the honest answer is more than we would like.

9.4 Giving the procedure its legal instrument

Rung C instructs the model to apply section and chapter notes, which are the legally binding instrument in HTSUS classification, and then never shows it one. That is the same defect we identify in the deployed system (§12), and it means the B→C null may be a null about a crippled implementation of the intervention rather than about the intervention itself. We harvested the notes from the USITC reststop endpoints and verified the extraction against USITC's own chapter PDFs, recovering chapter notes for 56 of the 59 benchmark chapters and section notes for 43; the 16 gaps are absences in the HTSUS rather than retrieval failures. Of 1,872 extracted lines checked verbatim against the PDFs, 1,712 matched exactly and three remain unexplained, all rows of one country-quota table.

Two arms were run, paired item for item on batch composition verified identical across all 37 key files: rung C exactly, and rung C plus the verbatim section, chapter and additional U.S. notes for every chapter appearing in that item's shown candidates, deduplicated per batch. Notes are included whole or named as omitted; the sole abridgement is chapter 98, whose 251,532 characters of additional U.S. notes do not fit any prompt, and which is recorded in the manifest. Batch size is 5 rather than the ladder's 13 because the appendix runs to 150k characters per batch, and *both* arms were re-run at 5, so batch size is held constant across the comparison.

24 of the 37 planned batch pairs completed (120 items). Batches are formed after a deterministic shuffle, so this is a random subsample rather than a biased prefix, but it is a partial run and the design detects roughly 17 pp at 80% power.

The notes arm leads at every depth and no difference is significant. At the duty-setting line accuracy moves from 59.2% to 64.2%, a paired gain of +5.0 pp (95% CI (−0.6, 10.6), exact McNemar $p = 0.15$, discordant cells 3/9); at 6 digits it is +7.5 pp ($p = 0.02$) and at 10 digits +5.0 pp ($p = 0.18$). The 6-digit p is the only one below 0.05, and it does not survive Holm correction across the three depths tested ($3 \times 0.022 = 0.067$), so we report no significant effect here. The honest reading is that supplying the legal instrument does not rescue the structure intervention the way supplying legible candidate text did: the point estimate is positive and small, the interval excludes the +17.7 pp that untruncating the candidates bought (§9.3), and this design cannot separate a small real effect from none. That is a weaker claim than

Depth	Accuracy (%) C (no notes)	C +notes	Δ	95% CI	McNemar p	b/c
6-digit	67.5	75.0	+7.5	$(1.8, 13.2)$	0.0225	2/11
8-digit (duty-setting)	**59.2**	**64.2**	+5.0	$(-0.6, 10.6)$	0.1460	**3/9**
10-digit (full code)	41.7	46.7	+5.0	$(-1.0, 11.0)$	0.1796	4/10

Table 5: Rung C with and without the binding section and chapter notes, paired on the same items at the same batch size. The notes arm leads at every depth, and no difference reaches significance at this sample size.

we would like and a stronger one than the ladder alone supported, because the intervention has now been run in the form its proponents describe rather than in the crippled form we first implemented.

An effect we did not set out to measure, and which costs us. Running rung C at two batch sizes isolated items-per-prompt from everything else. On the 125 items scored in both, the same prompt scores 36.8% at 13 items per prompt and 60.0% at 5: +23.2 pp (95% CI $(14.3, 32.1)$, exact McNemar $p = 2 \times 10^{-6}$, discordant cells $5/34$). This is larger than any intervention in the ladder. It does not touch the paired contrasts, which all hold batch size fixed, and it does not touch P13, whose arms both run at 5. It does mean that the *absolute* accuracies reported throughout §9 are substantially a property of how many items shared a prompt, and should not be read as this model's standing on this task. We had already found a within-batch position effect and stated that batching could induce order effects; this quantifies it, and the quantity is large enough that we would not have chosen batch 13 had we measured it first.

9.5 Does the ladder shape depend on the model?

Every result above rests on one model, which was the sharpest limitation of the previous version of this work and is the objection we most wanted to close. We therefore replicated the key contrast (A direct → B +retrieval) on `claude-haiku-4-5` (hereafter Haiku 4.5), a far smaller model than the `claude-opus-5` (Opus 5) used everywhere else. The prompts were byte-identical, the items the same ($n = 91$, the first seven ladder batches), and the scoring the same paired exact McNemar (Table 11 (appendix), Table 11 (appendix)). Fourteen further audited trials; nothing else changed.

The weaker model gets nothing right closed book. At the duty-setting 8-digit line, Haiku 4.5 is correct on 0 of 91 items (0.0%, 95% interval $(0.0, 4.1)$), and on 17.6% of them with the same fifteen candidates in the prompt: +17.6 pp, $p = 0.00003$. The discordant cells are perfectly one-directional, $b = 0$ against $c = 16$. Every item retrieval fixed, it fixed outright, and it broke nothing: no item that the model got right without candidates was lost when candidates were added, because there were none to lose.

Most of what retrieval supplies is a vocabulary, not a judgment. The structural check of §13 makes the mechanism visible: closed book, 5.5% of the small model's 10-digit codes exist in the HTSUS at all; with candidates in the prompt, 95.6% do (on the same items the large model moves 50.5% → 80.2%). For a weak model retrieval is not an enhancement but essentially the entire capability, and the first thing it supplies is not reasoning but a set of strings that are real tariff lines: the same mechanism §13 isolates on the large model, magnified until it is the whole effect.

What is and is not significant here. The Opus 5 contrast on this subset is 26.4% → 34.1%, +7.7 pp, and is not significant ($p = 0.189$). That is a power result, not a contradiction: the subset is half the ladder sample, and the same contrast on the full $n = 181$ is +11.0 pp at $p = 0.0045$ (Table 3). We report the non-significant cell rather than expanding the replication until it moved, because the honest summary is the one the two rows make together: the direction agrees across a wide capability range, and the effect is *larger* on the weaker model. That is the pattern expected if retrieval supplies something the model lacks rather than amplifying something it has, and it is evidence against the reading that the ladder's shape reflects one model's particular competence.

The caveat, at full strength. Both models are Claude-family, so an architectural effect and a shared-lineage effect are not separable here. Nor does this lift the single-sample caveat: these are two more single draws, on 91 items, through the same inference path. What the pair does establish is that the effect is not an artifact of one model's capability level, and that it *steepens* at the weak end. What it cannot establish is what happens at the strong end, which is the question §9.6 answers, and the answer there is not the one this subsection anticipates.

9.6 Cross-vendor replication, and the limit of the retrieval claim

§9.5 closed the single-model objection and left the single-vendor one open. We therefore ran the full ladder on `gpt-5.5-2026-04-23`, a frontier model from a different vendor, against the same batch files sent byte for byte: 42 calls, all three rungs, every call logged with the model snapshot the API returned, the parameters it accepted, and its token counts. Two differences from the Claude arms are worth stating before the numbers, because both favour this arm's internal validity rather than ours. The prompts reached this model through an HTTP endpoint with *no tools bound at all*, which is a strictly stronger closed-book guarantee than a rule against tool use plus an audit. And the API rejected `temperature` while accepting a seed, so the sampling regime is documented per call rather than assumed.

Retrieval does nothing for the stronger model. GPT-5.5 answers 59.3% of the 91 items correctly at the duty-setting line closed book, 1.7 times what Opus 5 reaches with retrieval on the same items, and adding the same fifteen candidates

moves it to 60.4%: +1.1 pp, discordant cells 6/7, exact McNemar $p = 1.00$. The intervention that carries this paper's headline effect is, on this model, indistinguishable from doing nothing.

The structure null replicates, and that claim gets stronger. Adding the enforced GRI procedure moves GPT-5.5 from 60.4% to 54.9%, a change of −5.5 pp that is not significant ($p = 0.27$). Two vendors, opposite signs, neither detectable: the finding that a reasoning-side scaffold does not measurably help is the one result here that survives crossing vendors intact.

This is not memorization. A high closed-book score invites the explanation that the model was trained on these rulings. The model's snapshot is dated and CROSS-2026 straddles that date, so the check is available at no inference cost: on the 88 items published before the snapshot GPT-5.5 scores 65.9%, and on the 93 published after it scores 61.3%, a gap of 4.6 pp with Fisher exact $p = 0.54$. As with the Claude arms (§11), there is no detectable memorization premium. The model appears to know the schedule rather than the answers.

What this does to the paper's central claim. The claim that retrieval rather than reasoning is the binding constraint is *falsified as an unconditional statement*, and we say so plainly rather than restricting it quietly to the models that produced it. What the three arms support is a conditional claim, and a sharper one:

> The benefit of retrieval is governed by the gap between what the model already knows and what the retriever can supply. It is not a property of the task.

The evidence is monotone across a 59-point range of closed-book competence and two vendors: +17.6 pp for a model that answers nothing correctly on its own, +7.7 pp for one at 26.4%, and +1.1 pp for one at 59.3% (Table 12 (appendix)). The mechanism is arithmetic rather than mysterious. Our retriever puts the correct 8-digit line in front of the model for 30.9% of items (§9.2); a model that answers 59.3% of them unaided already knows more than that retriever can tell it. A retrieval scaffold helps while it is better than the model's parametric knowledge and stops helping when it is not, and the crossover sits inside the range of models available today rather than beyond it.

What this does not show. It does not show that retrieval is useless for strong models in general, only that *this* retriever is, at 30.9% gold-line presence. The obvious follow-up is the one §9.3 makes available and we did not run: B′ lifted Opus 5 by +17.7 pp purely by rendering the same evidence legibly, and whether a better retriever or a fuller rendering would move GPT-5.5 is answered in §10: the better retriever moves it to 54.7% and the fuller rendering to 64.1%, against 63.5% closed book. It does not overturn the metric result of §4, which is a property of the tariff schedule and holds whatever any model does. And three models on one task with one retriever and one benchmark is a pattern, not a law: the monotone relationship in Table 12 (appendix) rests on three points, and we fit no curve to it and quote no slope.

10 What Evidence Is Worth Depends on the Model

The experiments to this point vary one thing at a time against one model, and each answers a narrow question. Put together they answer a wider one, and the answer is not the one the retrieval-augmentation literature usually assumes.

We ran a grid. Two axes vary independently and everything else is held identical: the same 181 CROSS-2026 items in the same order, the same batch composition verified by md5 over every key file, the same task instructions, the same scoring. The first axis is what we initially called *model competence*, measured by closed-book accuracy and spanning three models from two vendors. §10.1 shows that at the batch size used here this axis is mostly a measure of robustness to prompt packing rather than of knowledge, and restates the finding accordingly; the grid below is what was measured, and that subsection is how it should be read. The second is *evidence quality*, in four settings: none; `bge-small-en-v1.5` top-15 rendered at 110 characters (rung B); `text-embedding-3-large` top-15 at the same 110 characters (rung B$^+$); and `bge-small` top-15 rendered untruncated (rung B′). B and B′ share a retriever and differ only in how much of it the model was shown, so that pair isolates *rendering*. B and B$^+$ share a rendering and differ only in the retriever, so that pair isolates *retrieval quality*. Rung C adds the GRI procedure to B's evidence.

The same intervention has opposite signs on different models. Replacing `bge-small` with `text-embedding-3-large`, and changing nothing else, is worth +33.1 pp to Opus 5 (32.6% → 65.7%, exact McNemar $p < 10^{-6}$, discordant cells 10/70), surviving family-wise correction decisively (§10.6). The identical substitution, measured against its own closed-book baseline, moves GPT-5.5 by −8.8 pp (63.5% → 54.7%, uncorrected $p = 0.040$, discordant cells 35/19); that arm does *not* survive correction (adjusted $p = 1.00$), so on this test alone we claim a *failure to help*, not a demonstrated harm. Two later sections revise both halves of this and should be read with it: §10.5 replicates the negative arm four times and upgrades it to harm, and §10.2 re-measures the positive arm with both models at their own ceilings, where +33.1 pp becomes +6.3 pp and stops being significant. The number here is what the grid measured; it is not what the intervention is worth.

The asymmetry is what matters and it does not depend on the weaker arm's significance. One retrieval intervention, one benchmark, one prompt: on one model it produces among the largest effects in this paper, and on another it produces nothing detectable, with a point estimate pointing the other way. Retrieval augmentation therefore has no model-independent value on this task, and a paper that measures it on a single model is not measuring a property of the method.

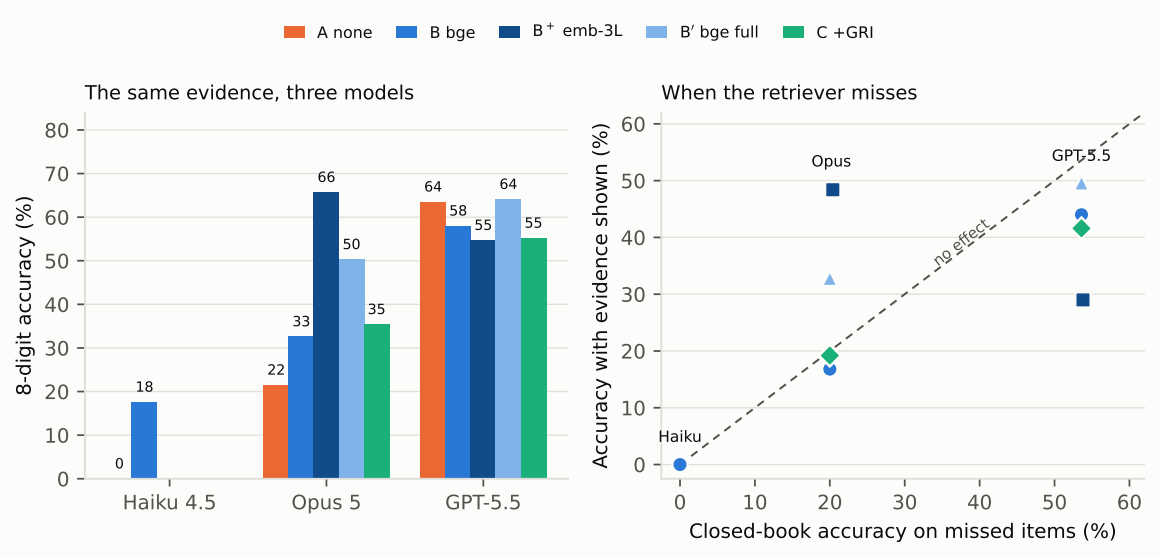

Figure 1: Left: the same four evidence settings applied to three models. The best configuration for Opus 5 is the worst kind of intervention for GPT-5.5. **Right:** the mechanism. Each point is one model and one rung, restricted to the items where *that rung's own retriever missed* the gold 8-digit line; the axes are the same model's closed-book accuracy on those items against its accuracy once the evidence was shown. Points above the dashed line mean the shown set helped even though it did not contain the answer; points below mean it displaced knowledge the model already had. The Opus cluster sits above the line, the GPT-5.5 cluster below it.

The mechanism: evidence acts on the prior, not only on the answer. Recall@k treats a retrieved set as valuable exactly insofar as it contains the answer. The split by retriever outcome shows that this is wrong in both directions. We split every rung on whether *its own* retriever surfaced the gold 8-digit line, and compare each half against the same model's closed-book accuracy on the same items, which is the control for what those items are worth with no evidence at all.

When the retriever succeeds, evidence helps every model: Opus 5 goes from 25.0% to 89.3% on those items under the best rendering, and GPT-5.5 from 85.7% to 96.4%. When the retriever *misses*, the sign depends on the model. Opus 5 gains 28.0 pp over its own closed-book rate on the items `text-embedding-3-large` failed to surface (20.4% $\rightarrow$ 48.4%): the shown set did not contain the answer but did narrow the space to the right neighbourhood. GPT-5.5 *loses* 24.8 pp on the same kind of items (53.8% $\rightarrow$ 29.0%). The evidence displaced knowledge it already had.

A mechanism we proposed, tested, and rejected. The stronger retriever hurts the stronger model *more* when it misses (-24.8 pp) than the weaker retriever does (-9.6 pp). Our first explanation was that a better retriever fails more *confidently*: on its misses, `text-embedding-3-large` returns a set whose chapter entropy is 0.96 bits against `bge-small`'s 1.32, so its wrong answers arrive as fifteen lines agreeing with one another rather than as a scatter. That story is testable at the item level, and it is wrong. Splitting the 93 missed items at the median of their own candidate-set entropy, the *concentrated* half costs GPT-5.5 19.6 pp and the *scattered* half costs it 29.8 pp: the opposite of the prediction. The entropy difference is also global rather than specific to failures (0.92 against 1.16 bits over all items), so it describes the retriever, not its misses. We report the refutation because the hypothesis was ours and it was attractive.

The mediator that does fit: deference. What separates the models is not the shape of the wrong evidence but how readily each model answers from inside it. On items where the retriever missed, GPT-5.5 still returns a code from within the shown set 57.0% of the time under B$^+$, against Opus 5's 28.0%. Within GPT-5.5 the relationship across all four evidence settings is monotone: deference of 13.6%, 26.4%, 28.0% and 57.0% corresponds to lifts of -4.0, -9.6, -12.0 and -24.8 pp. A model that defers to a candidate set which does not contain the answer pays the full price of the retrieval miss; a model that overrides it does not. The cost of deference is then roughly the deference rate times what the model would have got right unaided, which is small for Opus 5 (28% of 20.4%) and large for GPT-5.5 (57% of 53.8%), and that product is what the "Δ absent" column measures.

Two honest caveats. This is four points from one model, and it is correlational: we did not manipulate deference, and a model's willingness to override evidence is plausibly entangled with whatever else makes it accurate. And it does not explain everything, since Opus 5 shows no monotone relation of the same kind, which is expected if candidates that miss are still informative to a model that knew little, but means deference alone is not a complete account.

What this reorganizes. Three claims in the earlier sections survive unchanged and one does not. Retrieval carries the measurable gain *for models that lack the knowledge* (§9), and rendering the retrieved evidence legibly is worth more than adding a reasoning procedure to it (§9.3); both hold. The GRI scaffold is not detectably better than plain retrieval on any model or vendor (§9.6); that holds, and is the most portable negative result here. What does not survive is the unconditional reading of "retrieval is the binding constraint." The binding constraint is the gap between what the model knows and what the evidence supplies, and when that gap closes, the same scaffold that helped becomes a source of error.

10.1 The competence axis is largely a batching axis

Adversarial review put the sharpest possible question to the grid: every closed-book number in it is measured at 13 items per prompt, and §9.4 separately found rung C scoring 23.2 pp higher at 5 items per prompt. If the models differ in how much batching costs them, the axis the grid calls competence is measuring something else. We ran the control, and the objection is correct.

Closed book, each model scored on the items both of *its* arms covered, moving from 13 to 5 items per prompt is worth $+\mathbf{36.0}$ pp to Opus 5 (26.0% $\rightarrow$ 62.0% on $n = 100$, discordant cells $4/40$, exact McNemar $p = 1.7 \times 10^{-8}$) and $+\mathbf{0.6}$ pp to GPT-5.5 (63.5% $\rightarrow$ 64.1% on $n = 181$, $10/11$, $p = 1.00$). Opus 5 reads 26.0% rather than the 21.5% quoted elsewhere because the batch-5 arm covers 100 of the 181 items and the comparison is restricted to those; the GPT arm covers all 181. The gap between the two models is 37.5 pp at 13 items per prompt and 2.1 pp at 5.

The two models are not far apart in what they know about the tariff schedule. They are far apart in how much prompt packing costs them. Nearly the whole capability difference the grid was built around is an artifact of a design choice we made for throughput, and we would not have found it without being told where to look.

What this does and does not overturn. The central observation stands: both models were run on byte-identical prompts at the same batch size, so "the same evidence intervention is worth +33.1 pp to one model and harms the other" is a fact about two systems under identical conditions, not an artifact of unequal treatment. What fails is our *explanation*. We attributed the difference to how much each model already knew, and at matched batch size they know about the same amount. Closed-book accuracy at 13 items per prompt is not a measure of knowledge; it is a measure of knowledge *under load*.

The repair is a better-stated claim, and the data support it directly. What predicts whether evidence helps is not how much a model knows in the abstract but how far below its own ceiling it is operating when the evidence arrives. At 13 items per prompt Opus 5 scores 26.0% against its own 62.0% unimpaired, so it has a 36-point deficit for retrieved evidence to close, and evidence closes much of it. GPT-5.5 has essentially no deficit, so there is nothing for evidence to add and its errors can only subtract. On this reading the mechanism of §10 is unchanged, since deference converts a retrieval miss into a loss exactly when the model would otherwise have been right, and the induced deficit explains why Opus 5 had so much room to gain.

The cost of this to the rest of the paper. Every ladder number in §9 is measured at 13 items per prompt on a model that batching costs 36 points. Those contrasts remain internally valid, because all rungs share the batch size, but their absolute levels describe an impaired configuration, and the gains from retrieval reported there are gains against an impairment we introduced. We flag this wherever those numbers appear rather than restating them. §10.3 re-runs the ladder with this impairment and two others removed, and finds it reads 62.0%, 69.0% and 73.0% at the 8-digit line rather than 21.5%, 32.6% and 35.4%.

10.2 The grid at an unimpaired batch size, and what it costs us

§10.1 showed that packing 13 items into a prompt costs Opus 5 36 points of closed-book accuracy and GPT-5.5 nothing. The whole grid was measured at that batch size. A reviewer put the obvious question: was the large Opus gain from a better retriever knowledge the retriever supplied, or repair of an impairment we introduced for throughput? We re-ran the evidence conditions at 5 items per prompt, on batch composition verified identical across every arm.

The answer is that it was mostly repair. On 95 paired items with both models at their own ceilings, upgrading from `bge-small` to `text-embedding-3-large` is worth +6.3 pp to Opus 5 (62.1% $\rightarrow$ 68.4%, exact McNemar $p = 0.26$) and -7.4 pp to GPT-5.5 (62.1% $\rightarrow$ 54.7%, $p = 0.17$). The same contrast at 13 items per prompt was $+44.2$ pp and -8.8 pp.

Almost the entire headline gain was an artifact of our own design. We report the corrected figure rather than the one that made the better story: at a batch size where the model is not impaired, a production-grade retriever buys no benefit we can detect on this task, for either model.

Two things survive, and we are careful about how much weight they take. The *sign* still differs across models, which is the direction the headroom account predicts, but neither contrast is significant at $n = 95$ and we therefore treat the divergence as a pattern rather than a demonstrated effect. And the closed-book accuracies confirm the reading of §10.1: the two models are within 2.1 pp of each other unimpaired, so what the grid called competence was almost entirely robustness to prompt packing.

10.3 The ladder with every impairment removed

The structure null of §9 is the result a reviewer should trust least. Rung C instructs the model to apply section and chapter notes, which are the binding legal instrument, and the prompt never supplies them; its candidates are truncated at 110 characters, which §9.3 shows costs more than any intervention we added; and it runs at 13 items per prompt, which §10.1 shows costs this model 36 points. Three impairments, all of them ours, sit between the hypothesis and the test. Reporting a null under those conditions tests our implementation of structured legal reasoning, not structured legal reasoning.

So we ran the version we should have run first. Both arms receive the production-grade retriever, the same candidates rendered in full, the verbatim section, chapter and additional U.S. notes for every chapter in play, and 5 items per prompt. They differ in exactly one thing: whether the enforced GRI decision procedure is present. Both arms cover the whole

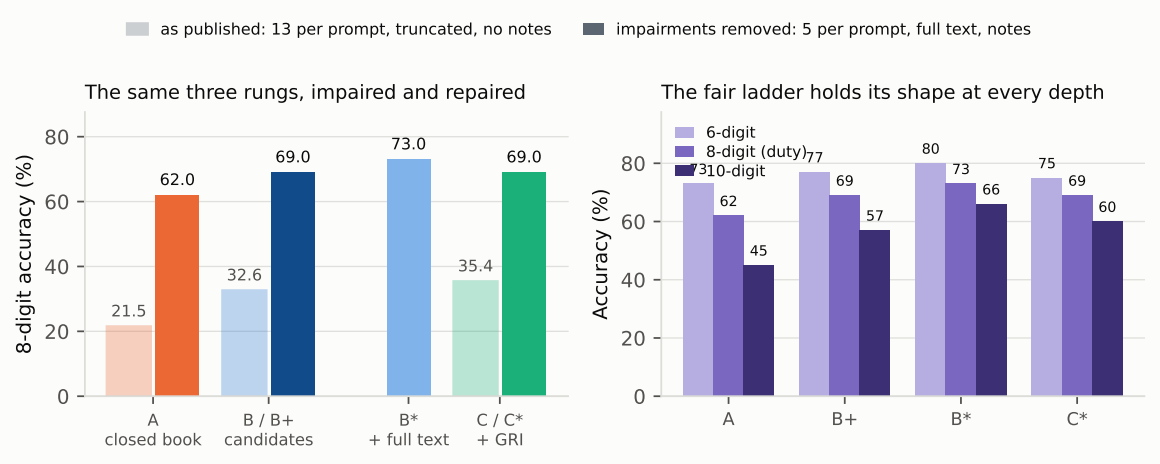

Figure 2: Left: the paper's published ladder (pale, $n = 181$, 13 items per prompt, candidates truncated, no notes) against the same three rungs with those impairments removed (solid, $n = 100$, 5 items per prompt, candidates in full, verbatim notes). Rung B* has no published counterpart because the published ladder never rendered its evidence in full. Right: the repaired ladder at all three depths, where its shape is the same one the 8-digit column shows. The two ladders are measured on different item counts, so the comparison is of levels within each, not a paired test across them.

benchmark, 181 paired items over 37 batches each, where an exact McNemar detects about 10 points at 80% power. This is the first structure test in the paper that is neither crippled nor underpowered.

Structure is negative at every depth and small everywhere. At the duty-setting 8-digit line it is -1.1 pp (69.1% $\rightarrow$ 68.0%, discordant cells 12/10, $p = 0.83$); at 6 digits -2.8 pp (9/4, $p = 0.27$); at 10 digits -2.2 pp (14/10, $p = 0.54$). The 8-digit interval is $(-6.2, +4.0)$ pp, which excludes any benefit larger than about four points. An interim reading at 50 items put the 8-digit delta at -6.0 pp; the full sample moves it to -1.1, and we report both because the difference between them is what 131 more items buys and is a fair measure of how much a 50-item panel should be trusted. The conclusion is now a properly powered null rather than a shrug: under the most favourable conditions we could construct, structure neither helps nor hurts. The tally across every configuration we measured it in reads the same way. It is $+2.8$ pp on Opus 5 at 13 items per prompt with truncated `bge` candidates ($p = 0.44$), the arm where the evidence was least legible and the only positive point estimate we have; -2.8 pp on GPT-5.5 in the same arm ($p = 0.38$); -14.9 pp against the same evidence rendered in full ($p = 10^{-4}$, surviving family-wise correction); -7.4 pp at 5 items per prompt with `text-embedding-3-large` candidates ($p = 0.17$); and -1.1 pp here, on the largest and cleanest of the five. Structural validity is unchanged (91.7% without the procedure, 92.3% with it), so the procedure is not buying legality either.

What the ladder looks like once it is not impaired. The four arms of Figure 2 share an item set and a batch size, so they can be read end to end. They cover the 100 items the two batch-5 arms of §10.2 reach, which is why this n is smaller than the 181 of the structure test above. Closed book scores 62.0% at the 8-digit line. Fifteen retrieved candidates truncated as the published ladder rendered them bring it to 69.0%; the same candidates in full, with the notes the procedure was always supposed to apply, bring it to **73.0**%, a gain over closed book of $+11.0$ pp that is significant (discordant cells 5/16, $p = 0.027$) and does not survive Holm correction across the paper's 56 tests (adjusted $p = 1.00$; §10.6). Adding the GRI procedure takes it back down to 69.0%. The retrieval claim of this paper therefore survives its own corrections, and the structure claim stays where it was: not detectably better than plain retrieval, now under conditions that no longer excuse it.

Every accuracy level in this paper describes an impaired system. The published ladder reads 21.5%, 32.6% and 35.4% at the 8-digit line. The same pipeline with our own three impairments removed reads 62.0%, 69.0% and 73.0%. The paired contrasts in §9 remain internally valid, since every rung there shares the impairments, but their *levels* are not this system's accuracy and should not be cited as a measure of what agentic tariff classification can do. We leave the original numbers in place rather than restating the paper around the better ones, because the gap between them is the paper's most transferable finding: the largest effects we measured were repairs of damage we had done ourselves, and we found none of them without adversarial review.

10.4 Deference is causal, and one sentence controls it

Everything above measures what happens when retrieval quality varies. The mechanism analysis suggested a different lever: what the prompt tells the model to *do* with retrieved evidence. That was correlational, so we manipulated it.

Candidate sets are constructed rather than observed, which fixes retrieval quality by fiat. In HIT the gold 8-digit line is guaranteed present, inserted for the 93 items where the retriever had missed it. In MISS it is guaranteed absent: every candidate sharing the gold 8-digit prefix is removed and the set backfilled from deeper ranks. We verified that 0 of 100 MISS items retain a gold-prefix candidate and 0 of 100 HIT items lack one. Crossed with this is a framing factor: the list is presented either as the answer space ("choose the best match from it") or as fallible suggestions ("the candidates are frequently wrong; if none is correct, ignore the list and answer from your own knowledge"). The causal prediction,

registered before the runs, is an *interaction*: framing should be inert under HIT and decisive under MISS.

It is.

	HIT (answer present)			MISS (answer absent)			
Model	choose	advisory	Δ	choose	advisory	Δ	interaction
Opus 5 ($n = 25$)	88.0	88.0	+0.0	0.0	72.0	**+72.0**	+72.0
GPT-5.5 ($n = 95$)	83.2	78.9	−4.2	0.0	58.9	**+58.9**	+63.1

Table 6: The deference manipulation. Framing barely moves accuracy when the evidence contains the answer and moves it enormously when it does not. Candidate sets are synthetic, so the absolute accuracies are not system performance; the quantity of interest is the framing difference within a construction, and the MISS floor of 0.0% is partly definitional, since a compliant answer cannot be correct when the gold line has been removed.

On GPT-5.5 ($n = 95$) framing is worth -4.2 pp under HIT and $+58.9$ pp under MISS, discordant cells $0/56$: an interaction of $+63.1$ pp. On Opus 5 ($n = 25$) it is $+0.0$ and $+72.0$ pp, discordant cells $0/18$, $p = 8 \times 10^{-6}$. The mechanism is visible in the deference rate itself. Told to choose from a list that cannot contain the right code, the models answer from inside it on 96% (Opus 5) and 100% (GPT-5.5) of items and score 0.0%. Told the same list is fallible, they leave it on 80% and 79% of items and recover to 72.0% and 58.9%, close to their own closed-book accuracy. The knowledge was never lost. The framing suppressed its use.

This corrects our own characterisation. §10 described GPT-5.5 as the more deferential model, on the observation that it answered from inside a wrong candidate set 57.0% of the time against Opus 5's 28.0%. Under explicit framing the two are indistinguishable: both defer completely when told to choose, and both largely stop when told the evidence is unreliable. Deference is therefore not a stable property of a model. It is a property of the instruction, with a model-specific component that appears only under an ambiguous one, and the prompt our ladder used was ambiguous. We had attributed to the model what belonged to our prompt.

What this is worth, and what it is not. The intervention costs one sentence and no retrieval, model or inference change, and on evidence that cannot contain the answer it is worth more than any other intervention we measured anywhere in this paper. That is the practical result, and the comparison is between measured effects: the other interventions were not tested under MISS, where a construction that removes the answer leaves them nothing to recover. It is bounded in three ways. The constructed conditions are synthetic and neither is a deployable retrieval regime; a real system does not know in advance whether its retrieval hit. The Opus arm is $n = 25$ and significant only because it is perfectly one-directional. And the advisory framing has a cost we did not measure here: telling a model to distrust its evidence should reduce accuracy when the evidence is right, and the -4.2 pp HIT cell on GPT-5.5 is a first, underpowered look at what that costs.

10.5 Repeat sampling, and what it does to the weak arm

Every cell elsewhere in this paper is a single sample, because temperature was not controllable on the sub-agent path. The cross-vendor path is scriptable, so we re-ran the two arms of the sign-flip contrast, rungs A and B^+ on GPT-5.5, three further times each with everything byte-identical but the seed. Four independent samples per cell.

The run-to-run spread is small. Closed book, the four runs score 63.5, 66.9, 65.7 and 66.9% (mean 65.75, SD 1.60 pp); with the production retriever they score 54.7, 53.0, 49.2 and 51.4% (mean 52.08, SD 2.34 pp). Individual predictions remain far less stable than the aggregate: across the four runs, 37.0% of items change their 8-digit answer at least once and 50.3% change their 10-digit answer, which reproduces on a second vendor the instability §E measured on the first.

The contrast is not noise, and it is larger than we first reported. Computed independently within each run, the A→B^+ delta is -8.8, -13.8, -16.6 and -15.5 pp: four out of four negative, mean -13.67 pp, which is 6.8 pooled run-to-run standard deviations from zero (one-sample $t = -8.0$ on the four run-level deltas, $df = 3$). Three of the four within-run exact McNemar tests return $p \leq 6.2 \times 10^{-4}$, and the strongest, $p = 3.6 \times 10^{-5}$, clears the family-wise threshold of §10.6 on its own.

This supersedes the caution we recorded there. On a single sample the negative arm did not survive correction, and we declined to claim harm; with four samples it is unambiguous, and the single-sample estimate of -8.8 pp was the most conservative of the four. We therefore do claim it: on this task, at this retriever quality, adding retrieved evidence *harms* the strongest model we tested. The replication is the evidence, not the original test.

The honest limit is that this design is available for one vendor only. The Claude arms still rest on single samples with one measured swing, so where a Claude contrast is small we continue to report it as undetected rather than absent.

10.6 Family-wise correction, and what it costs us

The paper reports 56 distinct paired hypothesis tests. Earlier drafts corrected within a digit depth, which no longer covers the family a reader actually sees, so we apply Holm's step-down procedure at $\alpha = 0.05$ across every reported test at once. Three defects in our own accounting were found by adversarial review and are fixed here: the script built the family from a key absent from one artifact, so no ladder-depth test entered it at all; it double-counted one contrast reported under two names; and it omitted the experiments of §§10.2–10.3 entirely, which would have corrected over the family we happened

to run first rather than the family the paper presents. The family below is the corrected one, and it is larger and therefore stricter than the one an earlier draft reported.

Eighteen tests are significant uncorrected and twelve survive: the two retriever upgrades on Opus 5 (+44.2 and +33.1 pp, adjusted $p < 10^{-4}$), the three rendering effects at 8, 10 and 6 digits (+17.7, +24.9 and +12.2 pp, adjusted $p < 10^{-4}$ to 0.0090), *both* framing effects under MISS (+58.9 pp on GPT-5.5, adjusted $p < 10^{-4}$; +72.0 pp on Opus 5, adjusted $p = 0.0004$), the batch-size effect (+23.2 pp, adjusted $p = 0.0001$), the retrieval effect on Haiku 4.5 (+17.6 pp, adjusted $p = 0.0015$), B′ beating rung C head to head (+14.9 pp, adjusted $p = 0.0048$), B^+ beating B′ (−15.5 pp, adjusted $p = 0.0141$), and the A→C ladder contrast (+13.8 pp, adjusted $p = 0.036$). The deference manipulation is the only causal experiment here and it survives on both models, which is the single strongest thing we can say about any result in this paper.

Six nominally significant results do not survive, and two of them matter.

- **The original A→B retrieval effect on Opus 5** (+11.0 pp, $p = 0.0045$) has adjusted $p = 0.190$. The result that motivated this entire line of work is not family-significant. It is not thereby refuted, and the much larger $A \to B^+$ effect on the same model survives comfortably, but the specific +11.0 pp figure should be read as suggestive rather than established.
- **The fair ladder's retrieval effect** (§10.3, +11.0 pp, $p = 0.027$) has adjusted $p = 1.00$. It is the cleanest retrieval estimate we have, measured with every impairment of ours removed, and inside a family of 56 a nominal p of 0.027 is not enough. We state it as significant uncorrected and not family-significant, and we do not use it to argue anything the surviving rendering and retriever-upgrade effects do not already carry.
- **The single-sample negative arm** (GPT-5.5 $A \to B^+$, −8.8 pp, $p = 0.040$) has adjusted $p = 1.00$. On that evidence alone we would claim only a failure to help. It is not the evidence we rely on: §10.5 re-runs both arms four times, finds the effect in every run with a mean of −13.7 pp at 6.8 run-to-run standard deviations, and that replication is what supports the claim of harm. A single test inside a large family is the weakest way to establish an effect; repeating the measurement is the strongest.

The other three losses are the two GPT-5.5 rendering contrasts (+6.1 and +9.4 pp) and the 6-digit chapter-notes effect, none of which carries an argument. Tables print uncorrected exact p values, because that is what each test computed; this subsection governs where the two disagree.

11 Is There a Memorization Premium?

CROSS-2026's date partition exists to make one experiment possible: if CROSS-derived accuracies are inflated because the rulings are in training data (HSCodeComp's stated reason for avoiding CROSS entirely (Yang et al., 2026)), then at matched difficulty a model should do better on older rulings than on rulings issued after its cutoff. We ran that experiment; the model does not.

Design. We harvested 2,504 additional CROSS rulings from 2022–2024, put them through the identical extraction and leakage-control pipeline (1,801 survived), and matched each modern item to a historical one on exact chapter, on exact 4-digit heading where possible (114 of 181 pairs; the remaining 67 matched at chapter level only), and on description length within ±35%. Both members of every pair were classified closed book by the same model and interleaved through the same shuffled batches, so that era is not confounded with batch position or session. We tested the paired outcomes with McNemar's exact test, Holm-corrected across the five depths. This experiment is independent of the ladder: an earlier draft reused its modern arm as the ladder's closed-book rung, which introduced the confound of §9.1; the numbers here come from the original run and are unaffected, because both of its arms were batched identically to each other.

Result: a null. At the duty-setting 8-digit line, 2022–2024 rulings score 21.0% and 2026 rulings 22.1%: a difference of −1.1 pp in the *wrong* direction, $p = 0.896$, $p_{\text{Holm}} = 1.000$, with 28 pairs where only the historical member was right against 30 where only the modern one was. No depth is significant, and the signs do not agree with each other. A secondary check points the same way. Overconfidence is essentially identical across eras (+25.2 pp historical, +23.6 pp modern), so the model is not even more confident on the rulings it had the opportunity to memorize.

How far the null goes, and where it stops. A null is worth reporting only if it is bounded. With 58 discordant pairs at the 8-digit level, the paired 95% interval is $[-9.4, +7.1]$ percentage points. That excludes a memorization premium larger than roughly 7 points and does not exclude one of 3–5, which is smaller than this design can see and larger than several deltas this literature reports between systems.

Two boundaries matter more than the headline. First, only 64 of the 181 modern items were issued on or after 2026-06-01, i.e. strictly after the model's stated May-2026 cutoff; for the remaining 117 the comparison is in-window against in-window, which is close to a design that guarantees a null. Second, that clean subset is small: it gives 25.0% against 25.0% with discordant cells 13 and 13, a difference of exactly zero. Simulating the same exact test at its observed discordant rate, however, puts its minimum detectable effect at 80% power at 23.0 percentage points. "Exactly 0.0 pp, $p = 1.000$" on that subset is therefore uninformative about any plausible effect size, and should not be read as the clinching evidence its arithmetic resembles. The informative statement is the interval on the full 181 pairs, and it carries the caveat that two thirds of those pairs are not a temporal control at all.

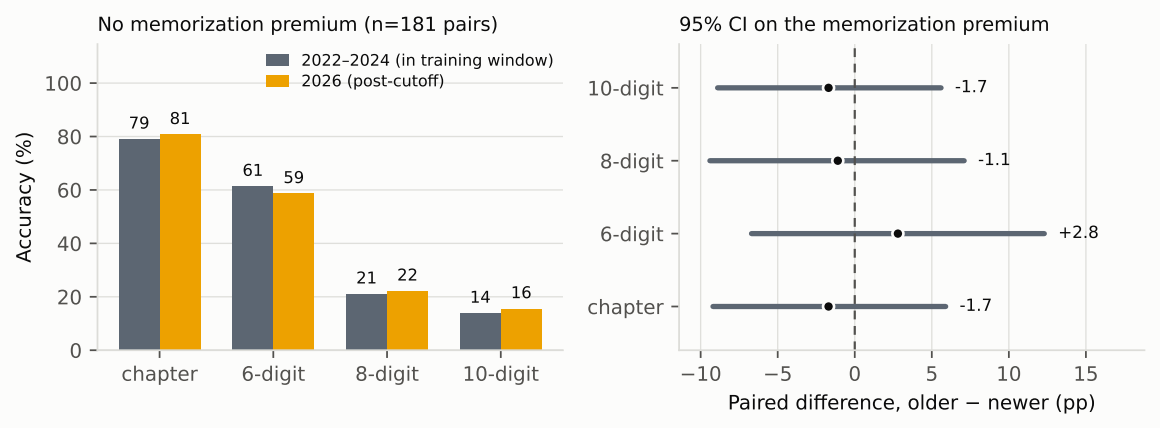

Figure 3: Left: accuracy by depth for the two eras on 181 difficulty-matched pairs, closed book. Issue era uses a neutral palette because it is orthogonal to the retrieval conditions of Tables 3 and 4. **Right:** the paired difference at each depth with its 95% interval. Every interval covers zero, and the intervals are wide enough that a small premium cannot be excluded.

This falsifies our own prediction F5, and it is good news for ATLAS. We predicted that accuracy on older, in-training-window rulings would materially exceed accuracy on post-cutoff rulings at matched difficulty, and we wanted to be right: a premium would have been the strongest available argument for CROSS-2026's date partition and against random splits over decades-spanning rulings corpora (Yuvraj and Devarakonda, 2025). It is not there. On this evidence the headline numbers of ATLAS-style CROSS-derived benchmarks are not inflated by a large memorization effect, and HSCodeComp's stated reason for avoiding CROSS is not supported by the one direct test we can run of it. That is a result reported against our own interest, and it removes much of the motivation for the date partition we built this benchmark around.

The remaining caveats are the obvious ones. This is one model, closed book; a memorization effect could appear with retrieval, where a remembered ruling might break a tie among candidates, and we did not test that configuration. Near-duplicate rulings mean a post-cutoff item can have a pre-cutoff twin (Zhang et al., 2025), which suppresses a real premium rather than manufacturing one, so it cuts toward our null being conservative. And a design partitioned on *documents* cannot see an effect operating on *labels*: matching on chapter, heading and length controls for the difficulty we can observe, not for whether the answer code was already familiar.

12 A Note on Vacuous Verification

The observation. We audited a deployed customs-classification pipeline (an unpublished TypeScript codebase) that reports "100% citation grounding" as a verification metric. It embeds the product description, retrieves the 50 nearest tariff-line chunks, renders them into the prompt, and forces a structured tool call returning a 10-digit code, an 8-digit code, the GRI rule applied, reasoning, citations, rejected alternatives, missing inputs, a precision level and a confidence. The grounding check is, in its entirety, `!candidateCodes.has(out.hts_code)` plus the analogous filter over the citation strings, where `candidateCodes` is the set of the 50 strings just shown to the model. It asks whether the answer is byte-identical to one of the fifty strings just displayed: a copy check, which cannot fail for a wrong-but-retrieved code, and wrong-but-retrieved is the dominant failure mode. The system's own diagnostic reports that for 20 of 37 investigated stubborn failures the correct code was in the top 50 and the model selected a different retrieved one. All 20 score 100% grounded. The check triggers at most one retry and blocks nothing; no caller reads the resulting warning field.

The constraint is not merely uninformative; it is costly. Through the retry, the predicate does not only report candidate-set membership, it enforces it. On the same 181 paired items, rejecting every out-of-set answer takes the retrieval rung from 32.6% to 17.1% at the duty-setting line, or 22.8% with one retry modeled optimistically (§C). Between a third and a half of what the system gets right, it gets right by leaving the list the grounding metric certifies. The metric is vacuous in the strict sense, and the behavior it rewards is expensive; those are separable defects, and the second is the costlier.

Three further findings bear on any claim that such a system performs structured legal reasoning. **(i)** Chapter and section notes never enter the prompt: they are used at index time, inside the embedding text, while the per-candidate rendering passes only the code and its description. Every prompt version from v2 instructs the model to quote the governing note verbatim; the model is never shown one, so any verbatim note in the output is recalled from parametric memory. **(ii)** Confidence is model-asserted against a prose rubric; the retrieval scores already in the trace are never used, and there is no calibration mapping. **(iii)** No sampling temperature is set anywhere in the codebase, so every previously reported delta is a single sample of unknown variance (cf. Yao et al., 2024).

A proposed definition. The observation generalizes past one codebase, and is the most transferable thing in this paper.

Definition (non-vacuous verification). A verification predicate V over a system's output is *non-vacuous*

with respect to a task if there exists an output that the system's own generation process can plausibly produce, that is wrong on the task, and for which V evaluates to false. A predicate that no wrong-but-retrieved answer can falsify is *vacuous*: it measures the pipeline's plumbing, not the determination.

Set membership in the retrieved candidate set is vacuous by construction, because retrieval supplies the vocabulary the answer is drawn from. So is a schema check over types and enumerations: a self-reported `gri_rule_applied` field drawn from a controlled vocabulary is a label, not a checked claim. So, arguably, is a same-model second pass over the same retrieval context, which inherits the blind spots that produced the error, consistent with Huang et al. (2024) and with HSCodeComp's finding that self-reflection and majority voting fail here (Yang et al., 2026). Non-vacuous predicates include: an entailment check that a cited note's text supports the claimed step, the target ALCE-style citation evaluation was designed for (Gao et al., 2023); a check that a cited precedent's holding does not contradict the predicted heading; cross-field invariants such as "`hts_code_8` is a prefix of `hts_code`"; and a counterfactual-dependency test that corrupts an evidence component and measures how often the answer changes (the intervention suggested by Turpin et al., 2023). The audited system implements none of these.

The definition applied to published work. An audit of one unpublished system is an anecdote, so we applied the definition to the verification mechanisms of seven published systems, reading each paper's full text (Table 7; per-system evidence and quotations in the released audit). The result is not the one this section's framing invites: under the definition as stated, *no system in the set is vacuous outright*. Two are non-vacuous, three partially so, one ships no verification predicate at all, and one turns on a deployer-supplied component we could not evaluate. What the audit finds is not a field of fake checks but a field whose checks are narrower than the prose around them, together with one specific check nobody performs: not one of the seven reports a measurement of whether its own evidence artifact constrains its own answer.

Two verdicts cut against us. Koomullil (2026) gate emission on a kernel-checked contradiction test and report catching all injected contradictions at a 6.4% false-block rate, a stronger verification story than the system we audited on every axis, and prior art we cite rather than improve on. Zhang et al. (2026b) identify the groundedness failure mode explicitly and build the output schema that would make it measurable; what is missing there is the measurement, not the concept. HSCodeComp supplies the strongest third-party support for our own claim, naming "Error but Valid" as the predominant agent error type and reporting that self-reflection does not fix it (Yang et al., 2026).

System	Verification mechanism	Verdict
ATLAS (Yuvraj and Devarakonda, 2025)	none at inference	no predicate
HSCodeComp (Yang et al., 2026)	expert oracle; gold match	non-vacuous
Zhang et al. (Zhang et al., 2026b)	schema; LLM-judged exclusion	partial
HSGraphAgent (Xia et al., 2026)	graph containment; LLM Redirect	partial
PCAA (Wang, 2026)	envelope schema; replay drift	partial
Koomullil (Koomullil, 2026)	Lean-checked emission gate	non-vacuous
Tagliabue & Greco (Tagliabue and Greco, 2025)	sandbox; human merge gate	partial†

Table 7: The definition applied to seven published systems, each read in full text. "Partial" means some component is non-vacuous and some is not. † the decisive component is a deployer-supplied predicate we could not evaluate. No system in the set is vacuous outright under the definition as stated; the picture changes under repair (i).

Four repairs the audit forced on the definition. Contact with real systems broke the definition in four places, each forced by a specific paper. **(i)** It did not distinguish an inference-time predicate from a gold-label metric: exact match against a held-out code is a predicate a wrong answer falsifies, so a benchmark scores as non-vacuous while being unusable as a deployment gate, where the gold label is precisely what is missing. We restrict V to predicates computable from the system's own inputs, retrieved context and output. This is what makes the audit bite: under it ATLAS has no V at all, and the runtime-available check in HSCodeComp reduces to code existence, which its own error taxonomy shows is vacuous against the dominant error. **(ii)** It was binary where systems are graded. We now grade by what discharges the predicate: *deterministic* (a decision procedure over an external artifact), *oracle-conditional* (a kernel-checked gate over a model-supplied signal), or *model-judged*. **(iii)** It was silent about the task over which "wrong" is quantified, which turns a true sentence about a system built for a different task into a misrepresentation of it; the task must be named, and a scope mismatch reported as scope rather than vacuity. **(iv)** Some predicates are parameterized by the deployer, so non-vacuity is a property of a deployment rather than of a system.

Under the repaired definition the audited system's check is still vacuous, now stated more sharply: an inference-time, deterministic predicate that no wrong-but-retrieved answer can falsify. The audit's more useful product is the repairs themselves. A definition that separates inference-time from gold-label predicates, grades by what discharges them, and carries a task-scope qualifier sorts systems rather than merely accusing them, which is the difference between a contribution and a complaint.

13 A Verification Predicate That Can Fail

The previous section is a critique; this one is its constructive counterpart. If the objection to `candidateCodes.has(...)` is that no wrong answer can falsify it, the reply must name a check that costs the same and *can*. Here is one:

whether the predicted code exists in the HTSUS at all. The check is a single hash lookup against the 19,949 published 10-digit lines, with no model call, no annotation and no retrieval context. Unlike set membership, it is a property of the answer rather than of the pipeline that produced it.

Closed book, most of the model's codes are not real lines. Applying the check to all three ladder rungs: direct and closed book, only 47.5% of the emitted 10-digit codes exist in the schedule, and only 55.8% are valid even at the 8-digit level: the model is not mostly picking the wrong line; it is mostly writing a string that is not a line. Adding retrieval raises 10-digit validity to 75.7% and structure holds it at 75.1%. A substantial part of what retrieval buys is that it stops the model inventing codes. The two closed-book arms of the contamination experiment (§11, a separate run) give exactly the same validity as each other, 46.4% in each era, which is a third independent line of evidence against a recall effect: a model reproducing memorized rulings should at least be reproducing real code strings more often for the era it could have memorized.

Validity is a strong error signal, and it yields the abstention that confidence did not. Conditioning on the check separates the predictions sharply: rung C is right at the 8-digit line 45.6% of the time when its code is a real line and 4.4% of the time when it is not, and the same split holds for the other two rungs. Used as a selection rule (answer when the code is valid, defer otherwise), it moves rung C from 35.4% at full coverage to 45.6% at 75.1% coverage (Table 13 (appendix)), i.e. +10.2 points of selective accuracy for a 24.9% abstention rate. Adding a self-reported confidence threshold on top takes it to 47.6% at 68.5% coverage. This is the behavior the retriever's score margin failed to produce (§7).

Both signals discriminate, including the one we wrongly wrote off. Selective accuracy at a chosen operating point confounds the signal with the threshold; the threshold-free measure is AUROC for predicting 8-digit correctness (Table 13 (appendix)). Structural validity scores 0.630 on rung B and 0.668 on rung C. Self-reported confidence scores 0.636, 0.683 and 0.626 on rungs A, B and C; on rung B it is the better of the two. Every interval excludes 0.5.

We had reported the opposite, arguing that model confidence was unusable for abstention because it is overconfident by 16–28 points (§D). That inferred a failure of discrimination from a failure of calibration, which does not follow: a usable abstention signal was in our own data while we were writing that we had not found one. The corrected statement is that this pipeline has two informative signals and no calibrated one: confidence needs a recalibration map rather than replacement, and the structural check needs no calibration because it is binary. It also settles prediction F4, which we had listed as untested.

What this does and does not show. The distinction between a verification predicate that can fail and one that cannot is not epistemic bookkeeping: here it is worth about ten accuracy points at three-quarters coverage for zero marginal compute, which is the strongest practical argument we can make for the definition in §12. Three limits bound it (L13). The gate is structural only: it asks whether a string is a line in the schedule, never whether that line describes the article, so it cannot catch the wrong-but-real code, the harder and more consequential error (P6b). Its size is a property of the system under test rather than a constant: a model emitting real codes 99% of the time would gain almost nothing, and §9.5 shows the other extreme at 5.5%. And the coverage points are single samples like everything else in §9.

14 Limitations

Corrections to our own analysis. Ten results in this paper are corrections to earlier versions of it (Table 8). We list them because they are what should calibrate a reader's trust in the rest, and because six of the ten surfaced only after someone asked where an unmeasured design choice was hiding. Four were found by recomputation from the committed per-item artifacts rather than by reading the manuscript, which is the argument for shipping the artifacts.

#	What we had claimed	What the control found	Found by
i	Rung A reused from the contamination arm was a fair closed-book baseline	It ran at 23.9 items per prompt against 12.9; re-run byte-identically, A→B moved +10.5 → +11.0 pp	review
ii	Self-reported confidence is unusable, being overconfident by 16–28 points	That infers a failure of ranking from one of level: AUROC 0.626–0.683, every interval above chance (§13)	recompute
iii	Dense retrieval beats BM25 by ≥20 pp; a memorization premium exists; the duty-weighted risk curve is monotone	+10.4 pp; −1.1 pp ($p = 0.90$); the curve reverses in the tail exactly as the accuracy curve does. All three withdrawn	recompute
iv	No benchmark item retains classification vocabulary	Scanning for CBP's determinative voice finds 40 of 698 (5.7%); §5 reports that instead	recompute
v	The 67.9% selection rate measures the model	It measures our 110-character rendering; untruncated it is 89.3%, and the rendering is worth +17.7 pp (§9.3)	review
vi	The grid's second axis is model competence	The two models are 37.5 pp apart at 13 items per prompt and 2.1 pp apart at 5 (§10.1)	review
vii	The retriever upgrade is worth +44.2 pp	At an unimpaired batch size, +6.3 pp ($p = 0.26$) on the same items; the ladder's levels fall out with it (§§10.2, 10.3)	review
viii	GPT-5.5 is the more deferential model	Under explicit framing both defer completely or barely at all; deference belongs to the instruction (§10.4)	review
ix	The Holm family covered every reported test	The script read an absent key, double-counted a contrast, and omitted the experiments run after review (§10.6)	recompute
x	The GRI procedure is at or below zero in every configuration	It is positive in one of five, by +2.8 pp; §10.3 gives the full tally	review

Table 8: Every correction this paper makes to its own earlier claims, with the control that forced it. "Review" means an adversarial reading named the unmeasured design choice; "recompute" means the committed artifacts contradicted the prose.

What follows are the objections that bear on the headline claims. §B continues with those that do not.

L1. Items are not independent, and the item-level intervals are optimistic by a measured amount. The 698 items carry only 389 distinct 10-digit labels; the 20 most frequent account for 193 items (27.7%), and one label (festive articles, 9505.90.6000) covers 37 near-identical rulings, which an item bootstrap resamples as 37 independent draws. An earlier draft listed the cluster bootstrap over 8-digit groups as not run; it has since been run (10,000 resamples of whole groups, seed 12345), and §G.1 reports it. The ladder intervals widen by about half (design effects 1.46 and 1.51 on the two deltas) and no headline claim changes status: A→B stays significant with a cluster interval of $(1.2, 22.2)$ pp, B→C stays non-significant at $(-5.1, 11.5)$ pp, and the contamination interval slightly *narrows*. Intervals reported elsewhere in this paper without a cluster counterpart should still be read as optimistic by a factor of roughly this size.

L2. The benchmark is small, unaudited, NY-only and conditioned on ambiguity. We harvested 91 HQ-track rulings (reconsiderations and the most contested determinations), and none survived extraction, because the extractor targets the NY letter structure while HQ rulings are issued as memoranda. CROSS-2026 is NY-only because our pipeline could not read the harder track, which biases it toward being easier in an unmeasured and fixable way. It has 698 items against ATLAS's 18,731, and labels are CBP's own determination with no re-adjudication, while we cite Zhang et al. (2026b), who found gold-label deviations in an expert-annotated benchmark. "CBP said so" is a strong answer for a binding ruling and a weak one for whether the label is recoverable from the text we retained. Rulings also exist because an importer asked, so the set is conditioned on contested goods, and we did not test whether those carry systematically larger duty spreads, which is the specific way that bias would interact with our duty-weighted headline.

L3. Answerability is bounded, not adjudicated, though the proxy shows no accuracy effect. CBP frequently decides a ruling on a submitted photograph, a physical sample, a laboratory analysis or technical literature, and our extraction keeps only ruling text. Scanning the original rulings for such references flags 314 of 698 items (45.0%), an *upper bound* on the unanswerable fraction, since a ruling can mention a photograph whose content is redundant with the prose. Splitting the ladder on our own flag finds nothing: at the 8-digit line rung C scores 32.4% on the 108 text-only-clean items against 39.7% on the 73 flagged ones (Fisher $p = 0.34$; rung A 20.4 against 23.3, rung B 30.6 against 35.6, both $p > 0.5$). Either the proxy has little validity or the ceiling does not bind at current accuracy; only a human audit separates them (P7), and none was performed.

L4. Three models, two vendors, two inference paths, and one sample per cell. The ladder is measured on `claude-opus-5`, its retrieval contrast replicated on `claude-haiku-4-5` (§9.5) and the full ladder re-run on `gpt-5.5` (§9.6). The cross-vendor arm closes the lineage objection and, in doing so, falsifies the unconditional form of our own retrieval claim. It introduces its own asymmetries, which we state rather than average away: the Claude arms ran through audited sub-agent sessions that read a batch file, the GPT arm through an HTTP endpoint with no tools bound, and temperature was uncontrollable on the first path and rejected by the model on the second. Both differences favour the GPT arm's internal validity, not ours. Every condition remains a single sample; we mitigate by pairing every comparison on identical items and by measuring the run-to-run swing (§E), but a fixed-seed harness would be strictly better. Prompts are ours rather than any published system's, so no number here reproduces anyone else's result. Three models is also not a curve: the monotone relationship in Table 12 (appendix) rests on three points and we fit nothing to it.

L5. The candidate list was rendered truncated, and the truncation cost more than anything else we varied. Each retrieved candidate entered rungs B and C as its code plus the first 110 characters of its ancestry-resolved legal text (`make_batches.py:201`). That text reads "heading > subheading > leaf", and the leaf distinguishes a line from its siblings; of the 2,715 candidate lines shown to the ladder, the cut removes the leaf entirely for 2,102 (77.4%; median candidate text 262 characters). This was flagged in an earlier draft as an uncontrolled confound and has since been run (§9.3): removing the truncation and changing nothing else lifts 8-digit accuracy from 32.6% to 50.3% and the selection rate from 67.9% to 89.3%. The limitation is therefore retired as an open question and recorded as a correction: rungs B and C, and every number derived from them, measure a pipeline whose evidence channel we had needlessly narrowed, and the $\approx 70\%$ selection figure was a property of our prompt rather than of the model. Rung C has still not been re-run against untruncated evidence, so the B→C null remains a null about structure *as we implemented it*.

L6. Part of §7 still rests on a weak baseline. The duty-weighted loss and chapter risk ranking are recomputed on the ladder's real predictions (§F) and both replicate. The *retrieval-margin* risk-coverage curves cannot be recomputed that way: they exist only for a retriever top-1 baseline and remain a property of a 6.0% classifier. The error-structure analysis of §6 (deepest correct prefix, confusion pairs, the chapter-98 attractor, the description-length tests) is computed on BM25 only.

15 Conclusion

The tariff schedule publishes its own loss function, and it assigns no loss to the last two digits of a 10-digit code: zero dispersion across 3,017 sibling groups and 25,651 pairs, zero column-level differences across 11,836 inheriting lines, and zero 10-digit references among 1,113 Chapter 99 trade-remedy citations. The field's headline metric therefore scores, in part, a distinction the tariff declines to make. Scored by consequence instead of string equality, 53.6% of our best classifier's 10-digit errors cost the importer nothing. That recomputation also exposed a property of economic metrics we had not anticipated: undefined for a code that does not exist, they silently condition on structural validity, and must be reported beside a validity rate or they will flatter systems that hallucinate.

The ablation that metric is meant to serve returned two results we did not want. The first is that the reasoning scaffold does not pay. An enforced GRI decision procedure is the component this literature and the system we audited invest most

heavily in, and its point estimate is positive in one of the five configurations we measured it in, by $+2.8$ pp ($p = 0.44$), in the arm where the evidence was least legible. Measured on all 181 items with the binding notes supplied, the candidates rendered in full and the batch size set where the model is unimpaired, it is -1.1 pp with a 95% interval of $(-6.2, +4.0)$: not a shrug about power, but a null we can defend. What pays is retrieval, and making the retrieved evidence legible pays more than reasoning over it: untruncating the same candidates is worth $+17.7$ pp and beats the GRI procedure head to head by $+14.9$ pp.

The second is about us. Three controls we ran only after adversarial review each removed a large part of an effect we had reported. Our own candidate rendering cost 17.7 points; our own prompt batching cost one model 36.0 points and the other 0.6, collapsing a 37.5-point apparent capability gap to 2.1; and with both repaired, a retriever upgrade worth $+44.2$ pp is worth $+6.3$ pp ($p = 0.26$). Every accuracy level in our published ladder describes a pipeline we had degraded in three ways: repaired, the same rungs read 62.0%, 69.0% and 73.0% rather than 21.5%, 32.6% and 35.4%. Under Holm correction over all 56 tests we report, twelve survive and six do not, including the $+11.0$ pp effect that motivated this work.

What survives crossing models is that retrieval augmentation has no model-independent value. The same retriever upgrade helps one model and harms another by -13.7 pp, replicated over k=4 seeds at 6.8 run-to-run standard deviations. Recall@k does not summarize a retriever, because it prices only the case where the answer is present and says nothing about whether the misses are informative or misleading. And what decides the sign is not the evidence but the instruction. Fixing retrieval quality by construction and crossing it with framing produces an interaction of $+63.1$ and $+72.0$ points, the only two causal estimates here and both family-significant: told to choose from a list that cannot contain the answer, models comply and score zero; told the same list is unreliable, they leave it and recover to near their closed-book accuracy. The knowledge was never lost. One sentence suppressed its use, and we had attributed to the model what belonged to our prompt.

We hand over a 698-item, date-partitioned, leakage-audited benchmark with a duty rate on every item; a paired ablation with its confounds removed and its noise floor measured rather than assumed; a causal demonstration that what a prompt says about its evidence matters more than the evidence; and a definition, that a verification predicate is non-vacuous only if a wrong-but-retrieved answer can falsify it, with a free check that satisfies it and buys $+10.2$ pp of selective accuracy. The largest gaps are a human answerability audit and the fact that no retrieval number here sits on the retriever the audited system deploys.

The lesson we would draw from the shape of this paper is larger than any result in it. The three biggest effects we measured were not properties of the task, the models or the method. They were damage we had done to our own pipeline, and each surfaced only because someone asked where an unmeasured design choice was hiding. A pipeline paper that reports no control of this kind has not shown that its controls came back clean; it has shown that it did not run them.

Reproducibility. Every non-model number here is produced by deterministic code from two public sources, the USITC HTSUS release (2026 Revision 16, retrieved 2026-08-19 through the `reststop` endpoints; both are live and unversioned in the URL, so we pin the revision here and commit the raw payloads with a SHA-256 per record) and CBP's CROSS rulings endpoint, with no GPU and no paid service required. The schedule, benchmark, retrieval and metric results additionally require no API key of any kind; the two commercial components, `text-embedding-3-large` retrieval and the `gpt-5.5` arm, do. The bootstraps and the power simulations are seeded (12345), the batch shuffling is seeded (20260819), and every table is generated directly from the committed JSON artifacts rather than transcribed. The model results require dispatching the generated batch prompts to a model session and saving the returned JSON, which is the one step that is not reproducible without inference access; the prompts, the opaque-id key files, the raw responses of both models, the tool-use audit and the scoring code are all committed, so the scoring is reproducible from our transcripts even where the generation is not. The post-hoc audits reported here (candidate rendering, post-cutoff coverage, the answerability split, residual evaluative language, minimum detectable effects and the retry-modeled constraint) are a single script over those artifacts and require no inference at all. A further script reads the manuscript's headline values back out of the artifacts and fails if any has drifted, so the prose is checked against the data on every rebuild rather than once by hand.

Appendix

The material below supports results stated in the body. It is separated rather than cut: every number the body cites is derived here, and every artifact remains committed.

A Supporting tables and figures

These exhibits are cited from the body. They are collected here so the argument is not interrupted by material that supports rather than carries it.

Construction step	Dropped	Remaining
Rulings harvested from CBP CROSS (NY track)	—	1,479
− not a classification ruling	358	1,121
− more than one article classified (cannot align description to code)	213	908
− description not cleanly separable from CBP's reasoning	145	763
− code absent from the current HTSUS revision	38	725
− code not a 10-digit line	27	**698**

Table 9: Subtractive construction of CROSS-2026. Every drop is recorded; no item was excluded for being hard. Separately, 91 harvested HQ-track rulings did not survive extraction at all; see L2.

Rung	gold line **present** ($n = 56$, 30.9%)	gold line **absent** ($n = 125$)
A direct	25.0	20.0
B +retrieval	**67.9**	16.8
C +structure	**71.4**	19.2

Rung	Stated conf.	Accuracy	Over-conf.
A direct	49.2%	21.5%	+27.7
B +retrieval	48.1%	32.6%	+15.6
C +structure	55.3%	35.4%	+19.9

Table 10: Left: 8-digit accuracy (%) conditioned on whether the correct line was among the 15 shown. Given the right candidate the model takes it about 70% of the time; without it, every rung collapses to roughly its closed-book rate. Both rungs render each candidate truncated to 110 characters; §9.3 shows that removing the truncation raises the present-column figure to 89.3%, so the values here are lower bounds on selection. **Right:** mean self-reported confidence against realized accuracy. Every rung is overconfident, and the structured rung is the most confident of the three.

	8-digit accuracy (%), $n = 91$ paired items						Code exists in HTSUS (%)	
Model	A direct	B +retr.	Δ	95% CI	McNemar p	b/c	A	B
`claude-opus-5`	26.4	34.1	+7.7	$(-2.1, 17.4)$	0.189	7/14	50.5	80.2
`claude-haiku-4-5`	**0.0**	**17.6**	+17.6	$(9.8, 25.4)$	3×10^{-5}	**0/16**	**5.5**	**95.6**

Table 11: Cross-model replication of the retrieval contrast, byte-identical prompts, $n = 91$ paired items, exact McNemar. b/c are the discordant counts (A-only correct / B-only correct). The right-hand columns give the share of emitted 10-digit codes that exist in the HTSUS (§13). The smaller model answers nothing correctly closed book and emits a real tariff line 5.5% of the time; retrieval moves both. The Opus row is not significant on this half-sized subset; the same contrast at $n = 181$ is +11.0 pp, $p = 0.0045$.

	8-digit accuracy (%), $n = 91$ paired items						Code exists (%)	
Model	A direct	B +retr.	Δ	95% CI	McNemar p	b/c	A	B
Haiku 4.5	0.0	17.6	+17.6	$(9.8, 25.4)$	3×10^{-5}	0/16	5.5	95.6
Opus 5	26.4	34.1	+7.7	$(-2.1, 17.4)$	0.19	7/14	50.5	80.2
GPT-5.5	59.3	60.4	+1.1	$(-6.7, 8.9)$	1.00	6/7	84.6	91.2

Table 12: The retrieval contrast across three models and two vendors, on the same 91 items with byte-identical prompts, ordered by closed-book accuracy. b/c are the discordant counts. The gain from retrieval falls monotonically as the model's own closed-book accuracy rises, and reaches zero.

Signal (AUROC for 8-digit correctness)	A direct	B +retrieval	C +structure
Self-reported confidence	0.636 (0.541–0.732)	**0.683 (0.598–0.759)**	0.626 (0.540–0.703)
Structural code validity	—	0.630 (0.576–0.680)	**0.668 (0.618–0.715)**

Selection rule	Coverage (%)	n answered	8-digit acc. (%)	95% CI
C, no abstention	100.0	181	35.4	(28.8, 42.6)
C, abstain on invalid code	**75.1**	**136**	**45.6**	**(37.5, 54.0)**
C, + confidence ≥ 0.5	68.5	124	47.6	(39.0, 56.3)
B, no abstention	100.0	181	32.6	(26.2, 39.7)
B, abstain on invalid code	75.7	137	40.1	(32.3, 48.5)

Table 13: Top: threshold-free discrimination, $n = 181$ per rung, percentile bootstrap intervals; AUROC = 0.5 is chance and every interval excludes it. The structural-validity row is reported for the two retrieval rungs, on which the selective-prediction analysis was run. **Bottom:** selective prediction under the structural-validity gate. Its intervals are computed in a separate analysis, so full-coverage figures differ from Table 3 in the last tenth of a point; the point estimates are identical. We claim only the rows shown: below roughly 35% coverage the curve is noisy and its intervals overlap everything.

B Further limitations

These bear on scope, provenance and statistical practice rather than on the headline claims, and continue the list of §14.

L7. The 8-digit claim is scoped to ad valorem tariff treatment, and is not weighted by trade. Restated from §4.3 because it is a correctness issue rather than a framing one: AD/CVD scope orders, agency admissibility flags (some operating at the 10-digit line), quota and country-of-origin marking are outside the HTSUS and untested, and 8.8% of lines carry rates we cannot reduce to a percentage. We also measure rate dispersion across schedule lines, not realized revenue. The trade-weighting half of this limitation has since been addressed and is reported in §G.2: the 8-digit invariance result cannot be changed by any weighting, and where weights are obtainable they move the economic argument in our favor, which we flag because it is the convenient direction. What remains untested is realized revenue at the 8- and 10-digit line, where the public statistics are behind an API key we do not hold, so the share of *import value* entering on lines we cannot reduce to a percentage is still unmeasured.

L8. Two retrievers, not the audited system's own. An earlier draft listed the production retriever as unmeasured. §10 now reports a second, production-grade retriever (`text-embedding-3-large`) alongside `bge-small-en-v1.5` and BM25, and the retriever turns out to matter more than any other variable we manipulated. What remains untested is `voyage-3-large` specifically, which the audited system uses and for which no key was available; and two retrievers do not make a curve, so the retriever axis of Figure 1 has two informative points, not a trend.

L9. The vacuous-verification audit reads papers, not systems. The definition is applied to seven published systems (§12), but from their papers rather than from their code, so a verdict reflects what a system documents rather than what it runs. Two verdicts rest on appendix material and one on a deployer-supplied predicate we could not evaluate.

L10. Statistical practice, denominators, and power. Comparisons use McNemar's exact test on paired outcomes with Holm's step-down correction, applied within each digit depth in the per-section tables and across the whole family of 56 tests in §10.6, which governs where the two disagree, and bootstrap intervals with a fixed seed (see L1). We did not run a prospective power analysis, and we did not correct jointly across the five depths of Table 3 and the three of Table 3, even though the prose does interpret non-primary depths, a wider family than the correction we applied. We also report no test of the difference *between* the two abstention signals, so "confidence beats validity on rung B" is an ordering of overlapping point estimates. Denominators differ by analysis and are stated inline: accuracy over $n = 698$, duty-weighted figures over $n = 636$, the "economically free" contrast over the 617 duty-scorable predictions wrong at 10 digits, and the ladder's duty-weighted rows over 86/136/135 (L13). On power: at $n = 181$ the paired test resolves the $+11.0$ pp retrieval effect comfortably and cannot resolve the $+2.8$ pp structure delta, whose 95% interval is $(-2.8, +8.4)$; simulating the same exact test at the observed discordant rate puts the minimum effect detectable at 80% power at 8.4 pp. Reporting that null as "structure does not help" would be a Type II error dressed as a finding; we say "not detectable" throughout, and the same bound applies to the negative A $\rightarrow$ B deltas at 6-digit and chapter. Separately, §E re-runs one condition once on 78 items, giving one realized run-to-run difference (1.2 pp) and not its distribution: no standard error, no ruling out larger swings on other rungs, and no claim that the 42.3% churn rate transfers ($k \geq 5$ repeats is the right instrument, P4b).

L11. Temporal holdout proves less than it looks like, and the contamination null is bounded. The benchmark spans 2026-02-11 to 2026-08-07 against a stated May-2026 cutoff, so only 259 of 698 items (37.1%) are strictly post-cutoff and the contamination experiment's modern arm contains 64 of 181. Near-duplicate pre-cutoff rulings, public knowledge of the schedule and the confound between schedule revision and ruling recency all apply (Zhang et al., 2025), and 38 items were dropped because their codes no longer exist in the current revision. Post-cutoff status is also not label novelty: a code first seen in 2022 can be the answer to a 2026 ruling, which a design partitioned on documents cannot see. The null excludes a memorization premium larger than about 7 points on 181 pairs, one model, closed book, but not one of 3–5 points, and does not test the retrieval-augmented setting where a remembered ruling might break a tie. The strictly post-cutoff subset ($n = 64$, $b = c = 13$), the only arm whose temporal control is unambiguous, has a minimum detectable effect at 80% power of 23.0 pp: "exactly 0.0 pp, $p = 1.000$" there is uninformative about plausible effect sizes.

L12. Prompt packing is a first-order variable and we measured it late. Items were batched for throughput, at 24 per prompt for the contamination arm, 13 for the ladder and 5 for the notes and control arms. §10.1 shows this is not a minor nuisance: closed book, going from 13 to 5 items per prompt is worth $+36.0$ pp to Opus 5 and $+0.6$ pp to GPT-5.5. Every ladder contrast holds batch size constant and so remains internally valid, but the ladder's absolute levels describe a configuration in which one model is operating 36 points below its own ceiling, and the gains retrieval shows there are gains against that self-inflicted deficit. §10.3 re-runs the ladder at 5 items per prompt with the truncation and the missing notes also repaired: the same three rungs read 62.0%, 69.0% and 73.0% at the 8-digit line. The structure contrast there covers all 181 items; the four-rung ladder is restricted to the 100 its closed-book and B$^+$ arms reach.

L13. The abstention signals are partial, and every duty-weighted number is conditioned on structural validity. The validity gate asks whether an emitted code exists, never whether it describes the article, so it cannot flag the wrong-but-real code (P6b), and its $+10.2$ pp is partly a property of a model whose closed-book validity is 47.5%. We report no recalibrated confidence signal (one sample per item at $n = 181$ will not support a reliability diagram), so the paper shows a usable signal exists without delivering a calibrated one. Relatedly, a prediction is priceable only if its code resolves to a rate, which requires it to exist, so §F drops 95/45/46 of 181 predictions per rung. Two consequences we cannot remove: a system that hallucinates more looks *cheaper* under an economic metric, its worst outputs leaving the denominator instead of entering it at a high cost; and the three rungs' duty-weighted rows are computed on three different subpopulations, so those differences are unpaired and carry no significance test, unlike every comparison in §9.

C What hard-constraining the answer to the candidate set would cost

The deployed ClearDeclare classifier validates `candidateCodes.has(hts_code)` and retries when the emitted code is not in the retrieved set (§12); Xia et al. (2026) take a stronger form of the same idea at decode time. Our rungs deliberately did not enforce it: the model was told it could answer outside the list, and did so on 54.1% (B) and 52.5% (C) of items. That makes the counterfactual computable (Table 14).

Rung	Unconstrained	Hard-constrained	Correct answers destroyed
A direct	21.5%	1.7%	36 (92.3%)
B +retrieval	**32.6%**	**17.1%**	**28 (47.5%)**
C +structure	35.4%	21.0%	26 (40.6%)

Table 14: What enforcing candidate-set membership would have cost, on the same 181 paired items, under the pessimistic assumption that a rejected answer is never recovered. "Correct answers destroyed" counts predictions right at the 8-digit line that would have been rejected for falling outside the 15 shown candidates; the percentage is of that rung's correct answers. Rung A was never shown a list, so its row measures nothing about the architecture and is reference only.

At a shown-candidate recall of 30.9%, rejecting every out-of-set answer takes rung B from 32.6% to 17.1%, destroying 28 of its 59 correct answers (47.5%), and rung C from 35.4% to 21.0% (26 of 64, 40.6%).

This is a bound, not a measurement, and three assumptions push it the same way. It scores a rejected answer as permanently wrong, where the deployed system retries. Under an optimistic single-retry model, in which an out-of-set answer is replaced by an in-set attempt succeeding at the rate we observe for in-set answers when the gold line is present, 73.8% (B) and 79.2% (C), the constrained estimate for rung B rises from 17.1% to 22.8% and rung C to 24.5%; "nearly halves" becomes "costs roughly a third," and that is still an assumption rather than a measurement. Our prompt also *invited* out-of-set answers in as many words, so we elicited the behavior we then charge the constrained architecture for; a genuinely constrained condition would produce different behavior, not the same behavior with rejections applied. And it runs at k=15 where the audited system retrieves 50, a depth at which the same retriever's 8-digit recall over the benchmark is 42.3% (§6) against the 30.9% presence rate at k=15 here. What survives all three is the conditional and its sign: a constraint that discards out-of-set answers is safe when recall is high and expensive when it is not. At these recalls it is expensive, and the assumption that it is free is wrong. Nothing here locates the crossover.

It also complicates our own prediction F2, that end-to-end accuracy of an answer-constrained configuration is bounded by its retrieval recall. That bound holds only if the constraint is enforced: unconstrained, rung B's 32.6% slightly exceeds the 30.9% presence rate because the model recovers answers the retriever never offered. F2 was mis-stated rather than confirmed or falsified; the corrected statement is that recall bounds accuracy exactly to the degree that the architecture forbids leaving the list.

D Calibration

Every rung is badly overconfident (Table 10, right): mean self-reported confidence exceeds realized 8-digit accuracy by $+27.7$ pp (A), $+15.6$ pp (B) and $+19.9$ pp (C). The cross-rung pattern is the uncomfortable one: adding the structured GRI procedure raises stated confidence by 7.2 points while adding 2.8 points of accuracy, so structure buys more confidence than correctness. Rung C, however, elicits confidence under a different output schema from A and B, a confound we cannot remove; the within-rung gaps are the safer reading.

A distinction we ourselves got wrong, and correct in print. An earlier version of this paper concluded from those numbers that self-reported confidence is not a usable abstention signal. That inference was invalid: it conflated *calibration* (whether the numbers are right in absolute

terms) with *discrimination* (whether the score ranks correct answers above incorrect ones). A badly calibrated score can discriminate perfectly, and this one discriminates well enough to be useful: AUROC for 8-digit correctness is 0.636, 0.683 and 0.626 across the rungs, every interval excluding 0.5 (Table 13). Model confidence here is miscalibrated but informative; an abstention policy built on it inherits the bias in the threshold, not in the ordering. We flag the error rather than quietly fixing it because it is the category mistake this paper criticizes elsewhere: judging a signal by whether its numbers look right instead of by whether it can discriminate. We report mean confidence against mean accuracy, not a reliability diagram: ECE and Brier decomposition over bins need more than one sample per item at $n = 181$. We do not attempt the recalibration this result implies is worth doing.

For completeness, the GRI rules rung C reported relying on were GRI 1 (132 items), GRI 3(b) (39), GRI 2(a) (5), GRI 6 (3) and GRI 3(c) (2). Per §12 this field is a self-reported label rather than a checked claim, and we use it descriptively only.

E Repeat reliability: how much of this is noise?

Temperature and seed are not controllable through the available inference path (§8), so every condition above is a single sample. That is the sharpest objection to the whole ladder, and it is answerable by measurement rather than argument: we re-ran rung C on the first 78 items with byte-identical prompts and compared the two runs.

Individual predictions are unstable, and that is a finding in its own right. Given identical input the model returns a different 10-digit code 42.3% of the time, a different 8-digit line 34.6% of the time and a different chapter 6.4% of the time. The declared 10-digit code is the number that goes on an entry summary and creates liability under 19 U.S.C. §1592; on this evidence, re-running the same article through the same system on the same day changes that number on nearly half of items. No accuracy figure, ours or anyone's, communicates that, and it is the property a compliance officer would ask about first. It is also the concrete form of Yao et al. (2024)'s objection: a determination right once in k runs is not a determination.

Aggregate scores are far more stable, because the per-item churn is close to symmetric: 22 items were correct at 8 digits in both runs, 4 in run 1 only and 3 in run 2 only, so measured accuracy moved 33.3% → 32.1%, a swing of 1.2 pp. This is what licenses reading the ladder at all: +11.0 pp is roughly nine times the observed swing and is not plausibly a sampling artifact, while +2.8 pp is about twice it and cannot be separated from it by a single sample per condition. The caveat is severe and stated rather than rounded off: two runs give one observation of the run-to-run difference, not an estimate of its variance (L10). What this establishes is a floor on the noise, not its distribution.

F Duty-Weighted Evaluation on the Real Classifier

The duty-weighted analysis of §7 was computed on BM25 top-1, at 6.0% 8-digit accuracy, because that was the only classifier we had at the time. We recompute it on the ladder's committed per-item predictions, and keep the BM25 version rather than replacing it, because the contrast is itself the result.

A denominator warning that is itself a finding. A prediction can be priced only if both its true and its predicted code resolve to an ad valorem rate, and a predicted code resolves only if it exists. Of the 181 true codes 180 resolve, so the true label is essentially never the binding filter; the prediction is. Only 86, 136 and 135 of 181 predictions (rungs A, B, C) are duty-scorable, and the excluded counts of 95, 45 and 46 track the counts of predicted codes that are not real HTSUS lines almost exactly (95, 44, 45; §13). Duty-weighted evaluation is therefore undefined for hallucinated codes and silently conditions on structural validity: the accuracies here are validity-gated and belong to the subpopulation whose answers could be priced at all, not to the headline 21.5/32.6/35.4% of §9. This is a property of any duty-weighted metric, not of our implementation, and its consequence is that such a metric must be reported beside a validity rate or it will flatter systems that hallucinate by dropping their worst outputs from the denominator.

The metric finding is stronger on a better classifier. On BM25, 34.8% of the predictions counted as failures at 10 digits carried zero duty consequence; on rung C it is 53.6% (52 of 97). At the 8-digit line the figure is 29 of 74 (39.2%) against 32.4% for BM25, and the share rises monotonically along the ladder (37.9%, 52.8%, 53.6% at 10 digits). That is the direction §4 predicts: a better classifier lands nearer the right line, and a near miss inside the correct 8-digit subheading is free by construction. It also moves the wrong way for anyone keeping exact 10-digit accuracy as the headline, since the better the classifier, the larger the fraction of its "errors" the tariff declines to charge for. The less flattering half of the same table: mean duty error is 1.94, 1.94 and 1.82 pp for A, B and C with overlapping intervals, so the accuracy gains do not translate into a significant reduction in duty error on the scorable subpopulation. What does move is the share exactly right on rate, 58.1% → 63.2% → 66.7%.

Direction of error, which $0/1$ accuracy cannot see. Rung C under-declares on 29 items and over-declares on 16 (B: 27 and 23; A: 17 and 19), with mean under-declaration of 5.74 pp at rung C. These are not symmetric costs: under-declaration creates retroactive liability and penalty exposure under 19 U.S.C. §1592, while over-declaration is in principle recoverable and in practice usually not recovered. No accuracy metric distinguishes them. We do not specify the asymmetric loss function that would price the difference, since that needs assumptions about audit probability and penalty multipliers outside anything we measured; making the asymmetry visible is the contribution.

Ranking chapters two ways on rung C's predictions reproduces on a real classifier what §7 found on BM25: by total duty error ch. 61 leads at 71.4 pp across 7 scorable items and ch. 64 follows at 33.0 across 5, while by error count the leaders are ch. 83 and ch. 39, and neither duty leader leads by count. The chapters a team would prioritize from an error-count dashboard are not the chapters carrying the money. The cell counts are small enough that this is a direction, not a ranking anyone should act on.

What this does to prediction F4. Applying the validity-then-confidence rule of §13 on the duty axis, rung C's mean duty error falls from 1.82 pp at full coverage to 1.61 pp at 25% and reverses to 2.28 pp at 10%; rung B falls further and more smoothly, 1.94 → 1.51 → 1.34 → 0.91 pp down to 50% coverage, then rises to 1.28 pp at 25% and reverses to 2.12 pp at 10%. The duty-weighted curve is measurable on a real signal and improves over most of its range, but shows the same tail reversal the accuracy curve does. F4 predicted it would be *better behaved* than the accuracy curve; on this evidence it is differently behaved, and both are non-monotone in the tail where the subsamples are small. We record F4 as measured and not supported in its comparative form.

G Robustness: Clustered Resampling and Trade Weighting

Two objections apply to every number above rather than to any one of them: the items are not independent, and the schedule-level analysis counts every tariff line equally regardless of how much trade crosses it. Both were listed as unaddressed in an earlier draft. Both are addressed here.

G.1 Cluster-robust intervals

The bootstrap intervals reported so far resample items. Items sharing an 8-digit gold label are not independent draws, so those intervals are too narrow. We re-estimated every headline quantity with a cluster bootstrap that resamples whole 8-digit gold-label groups (10,000 resamples, seed 12345, percentile interval, ratio estimator); the 181 ladder items fall into 128 groups.

Intervals widen and no headline claim changes status, but the margins are thinner than the item-level analysis implies, and that is the honest summary. The median design effect across all quantities is 1.19 on the standard-error scale (range 0.88 to 1.60); on the two ladder deltas it is 1.46 and 1.51. A→B still excludes zero, at $(1.2, 22.2)$ pp against the item interval of $(3.9, 18.2)$, but its lower bound falls from +3.9 to +1.2 pp and the companion cluster bootstrap gives $p \approx 0.027$ where the item-level exact McNemar gave $p = 0.0045$. The dependence is real rather than an artifact: the intracluster correlation of the A→B difference is 0.38. B→C stays non-significant, moving from $(-2.8, 8.4)$ to $(-5.1, 11.5)$ pp, which also means the structure null is a weaker null than the item interval suggested, admitting effects up to +11.5 pp. A→C stays significant at $(2.8, 25.6)$, and the cross-model contrast stays significant at $(7.9, 28.7)$. The exact McNemar tests are unchanged and are not restated as cluster-robust: they are item-level tests.

Two diagnostics support reading these as real dependence. A placebo check that permutes cluster labels at random, preserving the cluster-size distribution while destroying genuine dependence, returns design effects of 1.01 to 1.06, so the estimator is close to unbiased and the observed 0.88

to 1.60 is within-group correlation rather than a consequence of resampling unequal groups. Separately, the contamination interval *narrows* under clustering, from $(-9.4, 7.1)$ to $(-8.1, 6.6)$ pp, because the paired difference has a negative intracluster correlation (-0.38): within an 8-digit group the era-matched pairs tend to disagree in opposite directions. A sensitivity analysis clustering on connected components over both members' labels (91 clusters) gives $(-7.7, 6.8)$. We report the narrowing because it is what the data say; it tightens an equivalence bound around a null that was already our claim, and is not a new result.

G.2 Trade-weighted duty dispersion

The dispersion analysis of §4 counts each sibling group once. A reviewer will ask whether it survives weighting by the trade that actually crosses each line. For the headline claim the question resolves without any data at all: every one of the 3,015 8-digit groups with two or more ad-valorem-resolvable lines has a within-group duty spread of exactly zero, and a weighted mean of zeros is zero under any non-negative weighting. The 8-digit invariance result is therefore not merely robust to trade weighting, it is invariant to it deductively, and the trade data we could not obtain could not have changed it.

Below that line, weights matter and we obtained real ones. US import values by 6-digit HS subheading for 2025 come from the UN Comtrade public API (reporter USA, partner World, imports, annual), covering 5,469 subheadings and 95.6% of reported total US imports. Weighting by them, the mean within-6-digit duty difference falls from 1.681 pp counting each group once to 1.025 pp weighted (ratio 0.61); on the pooled-over-pairs aggregation the paper uses elsewhere the same comparison is 5.058 pp against 1.668 pp. And while 54.5% of 6-digit groups carry a single rate, those groups take 64.9% of matched import value. The 2024 vintage reproduces the weighted figure at 1.011 pp.

Trade weighting therefore moves the economic argument in our favor: duty variation below the 6-digit line is concentrated in subheadings that carry relatively little import value, so the claim that much classification error is economically free is stronger weighted than unweighted. We flag that explicitly because it is the convenient direction, and because the mechanism is visible and checkable rather than statistical: the largest US import subheadings are single-rate ones. Two things remain unmeasured. Comtrade publishes at 6 digits, so the 8- and 10-digit weights are still absent, and the US Census and USITC endpoints that carry them returned an explicit key error, logged with the HTTP response in the released artifact. In trade terms, 90.8% of import value sits in subheadings that are fully ad-valorem-resolvable, 3.4% in mixed ones and 5.9% in subheadings with no ad valorem rate at all, so the restriction stated in §4.3 costs less in value than in line count, though a mixed subheading cannot be split without the weights we lack.

G.3 How far the validity gate generalizes, and why it shrinks

The structural gate can only remove answers that are not real tariff lines, so its value is bounded above by how many such answers a configuration produces. That bound bites. Recomputing the same gate across every configuration we ran:

Model	Rung	Valid (%)	Ungated	Gated (Δ)
Opus 5	C	75.1	35.4	45.6 (+10.2)
Opus 5	B′	75.1	50.3	61.8 (+11.5)
Opus 5	B	75.7	32.6	40.1 (+7.5)
Opus 5	B^+	93.4	65.7	66.3 (+0.5)
GPT-5.5	B′	86.7	64.1	66.2 (+2.2)
GPT-5.5	B	90.6	58.0	59.8 (+1.7)
GPT-5.5	C	90.6	55.2	56.1 (+0.8)
GPT-5.5	B^+	94.5	54.7	53.8 (−0.9)

The +10.2 pp we report for rung C is real and is not representative. Across these eight configurations the gate is worth between −0.9 and +11.5 pp (the closed-book rungs, not shown, run higher still, to +18.0 pp on Opus 5, which only sharpens the point), and the ordering is almost exactly the ordering of the invalid-code rate: where a quarter of answers are not real lines it buys ten points, and where one answer in twenty is not real it buys nothing. Both the better retriever and the stronger model raise structural validity into the nineties, and the gate's headroom disappears with it.

We take two things from this. The gate remains the honest example the definition in §12 calls for, since it is a check a wrong answer can fail, and it costs nothing to compute. But it is a remedy for a failure mode that is already receding, and a paper proposing a cheap verification predicate should report the predicate's headroom on the strongest configuration available rather than on the weakest, which is the mistake we made in an earlier draft.

H Registered predictions, scored against ourselves

An earlier draft stated five predictions in advance so they could be checked against us. Four are now decided, and two of the four went against us in the direction that costs us most.

F1 (dense recall@50 for the correct 8-digit line exceeds BM25's 31.9% by ≥ 20 points) is *falsified*: the measured gain is +10.4 pp (§6). **F2** (accuracy of an answer-constrained configuration is at most its own recall) was mis-stated: it holds only where the constraint is enforced, and unconstrained rung B exceeds the presence rate by recovering answers the retriever never offered (§C). **F3** (a same-model critic does not significantly improve 8-digit accuracy) is untested; the critic rung was not run. **F4** (duty-weighted risk-coverage curves are better behaved than accuracy-based ones) is measured and not supported in its comparative form: the duty-weighted curve improves over most of its range and then reverses in the tail exactly as the accuracy curve does (§F), a differently behaved objective rather than a better behaved one. **F5** (accuracy on pre-2024 rulings materially exceeds accuracy on CROSS-2026 at matched difficulty) is *falsified* (§11). We said we would most like to be wrong about F1 and F5; we were wrong about both, and F5 removes much of the motivation for the date partition this benchmark is built around.

Five experiments listed as open in earlier drafts have since been run: the cross-vendor replication (§9.6), repeat reliability at k=4 on the cross-vendor arm (§10.5), the batch-size control (§10.1), the fair test of the structure rung (§10.3) and that test at full scale. What remains open, in priority order, is a third retriever, extending the fair ladder's closed-book and B^+ arms from 100 items to 181, a semantic entailment gate above the structural validity gate, a critic rung, and the human answerability audit, which needs adjudicators rather than trials: using a model for it would make the ground truth model-generated, which is the one thing this benchmark refuses to do.

I *

References

Avni, Dolev, Yagudaev, and Yung. Super-intelligence survival guide: Verification via proof-carrying output. IACR Cryptology ePrint Archive, Report 2026/994, 2026. URL https://eprint.iacr.org/2026/994. Author given names not recorded in the source used for this citation.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. Can GPT-3 perform statutory reasoning? In *Proceedings of the 19th International Conference on Artificial Intelligence and Law (ICAIL)*, 2023. URL https://arxiv.org/abs/2302.06100.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, et al. Why do multi-agent LLM systems fail? *arXiv preprint arXiv:2503.13657*, 2025. URL https://arxiv.org/abs/2503.13657.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, et al. 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)*, 2022. URL https://aclanthology.org/2022.acl-long.297/.

Simin Chen, Yiming Chen, Zexin Li, et al. Recent advances in large langauge model benchmarks against data contamination: From static to dynamic evaluation. *arXiv preprint arXiv:2502.17521*, 2025. URL https://arxiv.org/abs/2502.17521. Title reproduces the arXiv record verbatim, including the typo "Langauge".

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1), 2024. URL https://arxiv.org/abs/2401.01301.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL https://aclanthology.org/2023.emnlp-main.398/.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://arxiv.org/abs/1705.08500.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning (ICML)*, 2019. URL https://arxiv.org/abs/1901.09192.

Neel Guha, Julian Nyarko, Daniel E. Ho, et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2308.11462.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. A dataset for statutory reasoning in tax law entailment and question answering. In *Proceedings of the 2020 Natural Legal Language Processing (NLLP) Workshop*, 2020. URL https://arxiv.org/abs/2005.05257.

Jie Huang, Xinyun Chen, Swaroop Mishra, et al. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.01798.

Bryce Judy. Benchmarking harmonized tariff schedule classification models. *arXiv preprint arXiv:2412.14179*, 2024. URL https://arxiv.org/abs/2412.14179.

Saurav Kadavath, Tom Conerly, Amanda Askell, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

George Koomullil. Proof-carrying certificates for LLM pipelines: A trust-boundary architecture. *arXiv preprint arXiv:2605.16407*, 2026. URL https://arxiv.org/abs/2605.16407.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, et al. Conformal prediction with large language models for multi-choice question answering. In *ICML 2023 Workshop on Neural Conversational AI (TEACH)*, 2023. URL https://arxiv.org/abs/2305.18404.

Eunji Lee, Sundong Kim, Sihyun Kim, et al. Classification of goods using text descriptions with sentences retrieval. *arXiv preprint arXiv:2111.01663*, 2021. URL https://arxiv.org/abs/2111.01663.

Eunji Lee, Sihyeon Kim, Sundong Kim, Soyeon Jung, Heeja Kim, and Meeyoung Cha. Explainable product classification for customs. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–24, 2024. doi: 10.1145/3635158. Preprint arXiv:2311.10922.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://arxiv.org/abs/1711.06664.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. Hallucination-free? assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*, 2025. doi: 10.1111/jels.12413. URL https://arxiv.org/abs/2405.20362.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning (ICML)*, 2020. URL https://arxiv.org/abs/2006.01862.

George C. Necula. Proof-carrying code. In *Proceedings of the 24th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL '97)*, pages 106–119, Paris, France, 1997. ACM. doi: 10.1145/263699.263712.

Nicholas Pipitone and Ghita Houir Alami. LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*, 2024. URL https://arxiv.org/abs/2408.10343.

Junyu Ren. Evidence-grounded verified agentic reasoning: A path toward eliminating LLM hallucination in empirical inference via tool-attested kernel proofs. *arXiv preprint arXiv:2607.12650*, 2026. URL https://arxiv.org/abs/2607.12650.

Albert Sadowski and Jarosław A. Chudziak. On verifiable legal reasoning: A multi-agent framework with formalized knowledge representations. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM)*, 2025. URL https://arxiv.org/abs/2509.00710.

Manpreet Singh, Akshatha Srikantha, and Shyamal Lakhanpal. Cost-sensitive conformal prediction and human-in-the-loop abstention for imbalanced high-stakes decision support: A multi-domain benchmark. *arXiv preprint arXiv:2607.27143*, 2026. URL https://arxiv.org/abs/2607.27143.

Jacopo Tagliabue and Ciro Greco. Safe, untrusted, "proof-carrying" AI agents: toward the agentic lakehouse. *arXiv preprint arXiv:2510.09567*, 2025. URL https://arxiv.org/abs/2510.09567.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.04388.

Olivia Peiyu Wang, Sanna Wong-Toropainen, Daneshvar Amrollahi, et al. Know your limits: On the faithfulness of LLMs as solvers and autoformalizers in legal reasoning. *arXiv preprint arXiv:2606.16118*, 2026. URL https://arxiv.org/abs/2606.16118.

Zexun Wang. Proof-carrying agent actions: Model-agnostic runtime governance for heterogeneous agent systems. *arXiv preprint arXiv:2606.04104*, 2026. URL https://arxiv.org/abs/2606.04104.

Bingbing Wen, Jihan Yao, Shangbin Feng, et al. The art of refusal: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 2024. URL https://arxiv.org/abs/2407.18418. Also circulated as "Know Your Limits: A Survey of Abstention in Large Language Models".

Qiang Xia, Zijian Zhang, Ao Wang, Wenhan Wang, Xiangyu Wang, and Jian Li. HSGraphAgent: Knowledge-graph-guided large language models for harmonized system code classification. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 44761–44773, 2026. URL https://aclanthology.org/2026.acl-long.2072/.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024. URL https://arxiv.org/abs/2406.04244.

Yiqian Yang, Tian Lan, Qianghuai Jia, et al. HSCodeComp: A realistic and expert-level benchmark for deep search agents in hierarchical rule application. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026. URL https://arxiv.org/abs/2510.19631. Best Resource Paper. Preprint arXiv:2510.19631, 2025.

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

Pritish Yuvraj and Siva Devarakonda. ATLAS: Benchmarking and adapting LLMs for global trade via harmonized tariff code classification. *arXiv preprint arXiv:2509.18400*, 2025. URL https://arxiv.org/abs/2509.18400.

Terry Jingchen Zhang, Gopal Dev, Ning Wang, et al. Test of time: Rethinking temporal signal of benchmark contamination. *arXiv preprint arXiv:2509.00072*, 2025. URL https://arxiv.org/abs/2509.00072.

Yichi Zhang, Nabeel Seedat, Yinpeng Dong, Peng Cui, Jun Zhu, and Mihaela van de Schaar. Guideline-grounded evidence accumulation for high-stakes agent verification. *arXiv preprint arXiv:2603.02798*, 2026a. URL https://arxiv.org/abs/2603.02798.

Yu Zhang, Dongjiang Zhuang, Qu Zhou, et al. A deterministic agentic workflow for HS tariff classification: Multi-dimensional rule reasoning with interpretable decisions. *arXiv preprint arXiv:2605.14857*, 2026b. URL https://arxiv.org/abs/2605.14857.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. When does pretraining help? assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law (ICAIL)*, 2021. URL https://arxiv.org/abs/2104.08671.