# ATLAS: Benchmarking and Adapting LLMs for Global Trade via Harmonized Tariff Code Classification*

Pritish Yuvraj[1,*], Siva Devarakonda[1]

[1]Flexify.AI

**Abstract**

Accurate classification under the Harmonized Tariff Schedule (HTS) is a critical yet underexplored problem in global trade compliance, where errors can delay shipments and disrupt supply chains. We present ATLAS, the first benchmark and fine-tuned large language model for HTS code prediction, constructed from the U.S. Customs Rulings Online Search System (CROSS). The benchmark includes 18,731 legally grounded rulings spanning 2,992 unique codes, reformatted into reasoning-oriented prompts. Our fine-tuned ATLAS model (LLaMA-3.3-70B) achieves $40\%$ accuracy at the full 10-digit level and $57.5\%$ at the 6-digit level—improvements of $+15$ and $+27.5$ points over strong baselines—while being approximately $5\times$ cheaper to deploy. These results establish HTS classification as a rigorous benchmark for hierarchical reasoning, cost-efficient adaptation, and alignment in domain-specialized large language models. The dataset and model are publicly released to encourage further research on structured reasoning for real-world compliance tasks.

**Keywords**

large language models, hierarchical reasoning, benchmark, domain adaptation, fine-tuning, trade compliance, tariff classification, HTS code, LLaMA, structured prediction

## 1. Introduction

Every product imported into the global market must be assigned a Harmonized Tariff Schedule (HTS) code—a ten-digit identifier standardized by the World Customs Organization (WCO). The first six digits are harmonized globally, while the last four are country-specific; both are required for U.S. customs compliance [1].

The HTS is hierarchical: 22 sections expand into 99 chapters and thousands of subheadings, making tariff assignment a natural hierarchical learning problem. Six-digit accuracy reflects global consistency, while ten-digit accuracy measures U.S.-specific compliance.

Despite its centrality, classification remains a major bottleneck. The HTS spans over 17,000 pages, and recent U.S. policy changes mandate valid HTS codes for imports above $100. In 2025, postal operators such as India Post and Deutsche Post suspended deliveries to the U.S. due to missing or incorrect HTS codes [2, 3, 4], illustrating how fragile trade becomes without scalable automation.

Large language models (LLMs) offer a scalable alternative: their semantic reasoning and structured prediction capabilities suit fine-grained distinctions (e.g., semiconductor wafers vs. finished chips). Moreover, since the first six digits are globally harmonized, improvements in HTS classification can generalize worldwide while directly addressing U.S. compliance needs.

### 1.1. Contributions

We focus on the high-value semiconductor domain and present:

- The first open-source benchmark for HTS classification [5], derived from the U.S. Customs Rulings Online Search System (CROSS);

[0]1. HTS CROSS Rulings Dataset: https://huggingface.co/datasets/flexifyai/cross_rulings_hts_dataset_for_tariffs
2. Atlas LLM Model: https://huggingface.co/flexifyai/atlas-llama3.3-70b-hts-classification

*NORA'25: 1st Workshop on Knowledge Graphs & Agentic Systems Interplay co-located with NeurIPS, Dec.1, 2025, Mexico City, Mexico*

*Corresponding author.
pritish@flexify.ai (P. Yuvraj); siva@flexify.ai (S. Devarakonda)
https://www.pritishyuvraj.com/ (P. Yuvraj); https://tariffpro.flexify.ai/ (S. Devarakonda)

- A comprehensive evaluation of leading proprietary and open-source models, including GPT-5-Thinking, Gemini-2.5-Pro-Thinking, LLaMA-3.3-70B, DeepSeek-R1, and GPT-OSS-120B;
- The fine-tuned Atlas model [6] (LLaMA-3.3-70B), achieving $40\%$ 10-digit and $57.5\%$ 6-digit accuracy—substantially outperforming baselines—while being up to $8\times$ cheaper and fully self-hostable for privacy-sensitive deployments.

Together, these contributions position tariff code classification as a new benchmark for evaluating reasoning and adaptation in large language models. Traditional approaches to product classification typically rely on hand-crafted features (e.g., HS keyword rules, tariff-engine heuristics) or shallow models such as random forests and gradient-boosted trees trained on bag-of-words representations. While such models can work well for constrained taxonomies, they struggle to scale to thousands of fine-grained, legally nuanced classes and offer limited support for explanation, counterfactual analysis, or rapid adaptation to regulatory change. In contrast, large language models can ingest raw ruling text, reason over subtle distinctions (e.g., "partially fabricated" vs. "finished" wafers), and produce both a code and an accompanying rationale. A systematic comparison with non-LLM baselines is an important direction for future work, but in this paper we focus on establishing a strong LLM-based baseline and a reusable benchmark.

**Relevance to Knowledge Graphs and Agentic Systems.** Although this work focuses on supervised fine-tuning, HTS classification naturally interfaces with agentic systems and knowledge graphs. In practical deployments, a tariff-classification agent must (i) retrieve and ground its predictions in structured regulatory corpora (e.g., HTS knowledge graphs linking codes to legal notes and duty rates), (ii) orchestrate multi-step reasoning workflows (e.g., querying prior rulings, edge-case escalation, and audit trails), and (iii) interact with downstream customs-compliance pipelines. We position Atlas as a core reasoning component in such agentic systems: our benchmark and model provide a standardized, hierarchical decision layer that can be plugged into knowledge-graph–augmented retrieval and tool-using agents for end-to-end trade compliance.

## 2. Dataset

Our main contribution is the first large-scale dataset for Harmonized Tariff Schedule (HTS) classification, derived from the U.S. Customs Rulings Online Search System (CROSS) [7]. CROSS contains legally binding rulings by U.S. Customs and Border Protection (CBP) specifying the correct 10-digit HTS code for products. These rulings are authoritative yet dispersed across thousands of HTML pages, previously inaccessible for ML research.

**Example instance.** A typical CROSS ruling in our dataset describes, for example, a shipment of "12-inch silicon wafers that have undergone photolithographic patterning but are not yet diced or packaged." The corresponding HTS US code is `8542.90.0100`, which encodes (roughly) "electronic integrated circuits; other; wafers; unmounted." Our processed instance contains (i) a condensed product description, (ii) a reasoning-style explanation derived from the ruling (e.g., why certain headings are excluded), and (iii) the final 10-digit code.

### 2.1. Collection and Scope

We built an automated agent [8, 9, 10] to scrape CROSS and align each ruling with its official 10-digit HTS code from [1]. Focusing on semiconductor and manufacturing chapters, we obtained $18{,}731$ rulings covering $2{,}992$ unique codes. Frequent rulings highlight ambiguous or high-demand categories, while absent codes suggest stable ones. Table 1 lists representative chapters (the complete distribution is provided in Appendix A, Table 5).

**Table 1**
Sample distribution of CROSS rulings across major HTS chapters.

| Chapter | Codes | Rulings | Description |
|---|---|---|---|
| 39 | 264 | 2781 | Plastics and articles thereof |
| 61 | 163 | 3445 | Apparel, knitted or crocheted |
| 73 | 389 | 1749 | Articles of iron or steel |
| 84 | 801 | 3566 | Machinery and mechanical appliances |
| 85 | 293 | 1445 | Electrical machinery and equipment |
| 90 | 256 | 1499 | Optical and precision instruments |
| **All** | 2992 | 18731 | Combined total |

### 2.2. Transformation to Model-Ready Format

Raw rulings are verbose legal letters. We used GPT-4o-mini [11] for information extraction, converting each into a concise instruction–response pair containing (a) a product description, (b) reasoning trace, and (c) final HTS code. The complete prompt template is provided in Appendix B. This structure enforces reasoning-based prediction, aligning with chain-of-thought research [12].

### 2.3. Splits and Availability

We reserved 200 samples each for validation and testing, with the remaining 18,254 for training (Table 2). To ensure fair and representative evaluation despite the small test size, the 200 test samples were stratified across high-variance HTS chapters (e.g., 84, 85, and 90) to reflect the diversity and ambiguity observed in real-world tariff rulings. The dataset is publicly available on Hugging Face [5].

**Table 2**
CROSS dataset splits.

| | |
|---|---|
| Training | 18,254 |
| Validation | 200 |
| Test | 200 |

We chose relatively small validation and test splits (200 examples each) for two reasons. First, manual inspection and error analysis at the 10-digit level are time-consuming because each prediction must be checked against lengthy legal notes and chapter-specific carve-outs. Second, our goal was to maximize the training signal for ATLAS while still preserving a stratified and diverse test set that covers high-variance chapters (e.g., 84, 85, 90). In practice, we observe no evidence of overfitting on the validation set (see Section 3), but we regard scaling up the held-out set as important future work.

### 2.4. Discussion

HTS rulings demand fine-grained reasoning (e.g., partially vs. fully fabricated wafers) and hierarchical accuracy at 6- and 10-digit levels. Errors have direct compliance costs, making this dataset a realistic and impactful benchmark for evaluating structured reasoning in large language models.

## 3. Model Training

While several open-source large language models could, in principle, be adapted for tariff classification, we made a deliberate and principled choice to focus exclusively on **LLaMA-3.3-70B** [13]. Two factors motivated this decision. First, practical *budget constraints* made it infeasible to fine-tune multiple frontier models at scale. Second, LLaMA-3.3-70B is a dense architecture, making it both simpler to fine-tune and easier to deploy in inference settings compared to Mixture-of-Experts (MoE) architectures such as DeepSeek-R1 or GPT-OSS-120B. From a community perspective, providing a dense and reproducible

baseline lowers the entry barrier for downstream research: training and inference pipelines are easier to set up, memory usage is more predictable, and accuracy is less sensitive to expert routing heuristics.

### 3.1. Supervised Fine-Tuning Objective

We adapted LLaMA-3.3-70B to the CROSS dataset using supervised fine-tuning (SFT) [14, 15]. Each ruling was transformed into an input–output pair, where the input is a ruling-derived product description and the output is the correct HTS code along with a reasoning trace. This makes the task well aligned with the SFT paradigm, which minimizes the token-level negative log-likelihood of ground-truth outputs.

Formally, for an input sequence $x = (x_1, \ldots, x_n)$ and target sequence $y = (y_1, \ldots, y_m)$, the model with parameters $\theta$ defines conditional probabilities $p_\theta(y_t \mid x, y_{<t})$. The training loss is then:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{t=1}^{m} \log p_\theta(y_t \mid x, y_{<t}),$$

which corresponds to the standard negative log-likelihood objective.

### 3.2. Training Setup and Stability

Fine-tuning was performed for 5 epochs (approximately 1,400 steps) using the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay $= 0.1$, and a cosine learning-rate schedule initialized at $1 \times 10^{-7}$. To manage the high memory footprint of 70B-parameter models, we employed bf16 precision and gradient accumulation to simulate a batch size of 64 sequences. Training was distributed across 16 $\times$ A100-80GB GPUs using fully sharded data parallelism.

As shown in Figure 1, the training loss decreases sharply in the first 200 steps and then stabilizes near convergence, with no sign of overfitting on the validation set. We observed stable gradient norms and no catastrophic spikes in loss, suggesting that dense models like LLaMA-3.3-70B are well suited to small but domain-specific datasets when carefully regularized. This highlights that reproducible fine-tuning of frontier models is feasible even under modest compute budgets, provided that optimization choices are tuned to stability.

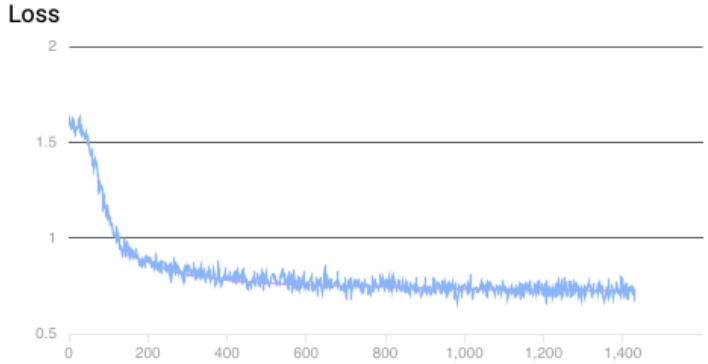


**Figure 1:** Training loss curve over 1,400 optimization steps. Rapid early improvement is followed by stable convergence.

### 3.3. Beyond Supervised Fine-Tuning: Reinforcement Learning

Although our experiments in this paper are limited to supervised fine-tuning, we view reinforcement learning (RL) as a promising extension rather than a component of the current Atlas training pipeline. Concretely, a lightweight and cost-effective starting point would be a **rule-based reward model**. For instance, we can define rewards as: 1 when the model correctly predicts the full 10-digit HTS US code, 0.6 when the first 6 digits (globally harmonized HS code) are correct, and $-1$ otherwise. Formally, for classification $\hat{y}$ and gold label $y$:

$$R(\hat{y}, y) = \begin{cases} 1, & \text{if } \hat{y}_{1:10} = y_{1:10} \\ 0.6, & \text{if } \hat{y}_{1:6} = y_{1:6} \\ -1, & \text{otherwise.} \end{cases}$$

Such a structured reward can be readily integrated into GRPO [16] or related policy-gradient methods such as PPO [17]. This approach would allow the model to explore reasoning trajectories that go beyond memorization, while keeping the reward landscape interpretable and inexpensive to compute. Importantly, this positions tariff code classification as a promising candidate for lightweight reinforcement learning research on high-stakes, domain-specific reasoning tasks. We leave the implementation and empirical evaluation of such RL schemes for HTS classification to future work.

### 3.4. Ablations and Future Work

While our study focused exclusively on LLaMA-3.3-70B, several ablation studies could provide deeper insights and further guide the community:

- **Model scale:** Evaluating smaller LLaMA variants (e.g., 8B or 3B) would clarify the tradeoff between accuracy, cost, and deployability on edge devices.
- **Retrieval augmentation:** Integrating retrieval over the 17,000-page HTS documents may reduce hallucinations and improve long-tail classification accuracy, complementing SFT.
- **Contrastive and hybrid objectives:** Beyond NLL, contrastive learning between closely related codes (e.g., semiconductor wafers vs. finished chips) may sharpen decision boundaries.
- **Direct Preference optimization:** Beyond NLL training, methods such as Direct Preference Optimization (DPO) [18] could leverage structured preferences over HTS classifications (e.g., preferring correct 10-digit codes over near-misses, or valid reasoning traces over hallucinated ones). This would allow the model to learn not just to imitate CROSS rulings but to actively steer away from incorrect classifications.
- **RL scaling studies:** Comparing rule-based GRPO with preference-based RLHF could quantify the cost–benefit tradeoffs of reinforcement learning at 70B scale.

These directions highlight that while Atlas establishes a strong dense-model baseline, HTS classification remains an open problem with substantial room for methodological innovation.

## 4. Results and Evaluation

We evaluate all models on a held-out test set of 200 CROSS rulings, predicting the correct 10-digit HTS US code per product. Because the classification is hierarchical, we report three metrics: (1) full 10-digit match, (2) partial 6-digit match (globally harmonized), and (3) average digits correct (0–10).

**Average digits correct.** Beyond exact 6- and 10-digit accuracy, we report *average digits correct*, which measures how many leading digits of the 10-digit HTS US code the model gets right on average. For a predicted code $\hat{y}$ and gold code $y$, we compute the longest matching prefix length $L(\hat{y}, y) \in \{0, 1, \ldots, 10\}$ and then average $L$ over the test set. This metric captures how close near-miss predictions are along the hierarchical tree: a model that consistently predicts the correct chapter and heading but misses fine-grained subheadings will have a high average prefix length even if its 10-digit accuracy is modest.

### 4.1. Accuracy at 10 and 6 Digits

Table 3 shows fully correct classifications. GPT-5-Thinking[1] achieves 25%, while **Atlas** attains **40%**, the highest among all models.. At the 6-digit level (Table 3), Atlas also leads with **57.5%**, slightly

[1]Model outputs and pricing were obtained from public API documentation and experiments conducted in September 2025.

above GPT-5's 55.5%. These results confirm that domain-specific fine-tuning improves both global and U.S.-specific accuracy.

| Model | 10-digit (%) | 6-digit (%) | Avg. Digits |
|---|---|---|---|
| GPT-5-Thinking | 25.0 | 55.5 | 5.61 |
| Gemini-2.5-Pro-Thinking | 13.5 | 31.0 | 2.92 |
| DeepSeek-R1 (05/28) | 2.5 | 26.5 | 3.24 |
| GPT-OSS-120B | 1.5 | 8.0 | 2.58 |
| LLaMA-3.3-70B | 2.1 | 20.7 | 3.31 |
| **Atlas (fine-tuned)** | **40.0** | **57.5** | **6.30** |

**Table 3**
Model accuracy across 10-digit, 6-digit, and average-digit metrics.

Figure 2 visualizes the relative advantage of Atlas, particularly for 10-digit U.S.-specific codes.

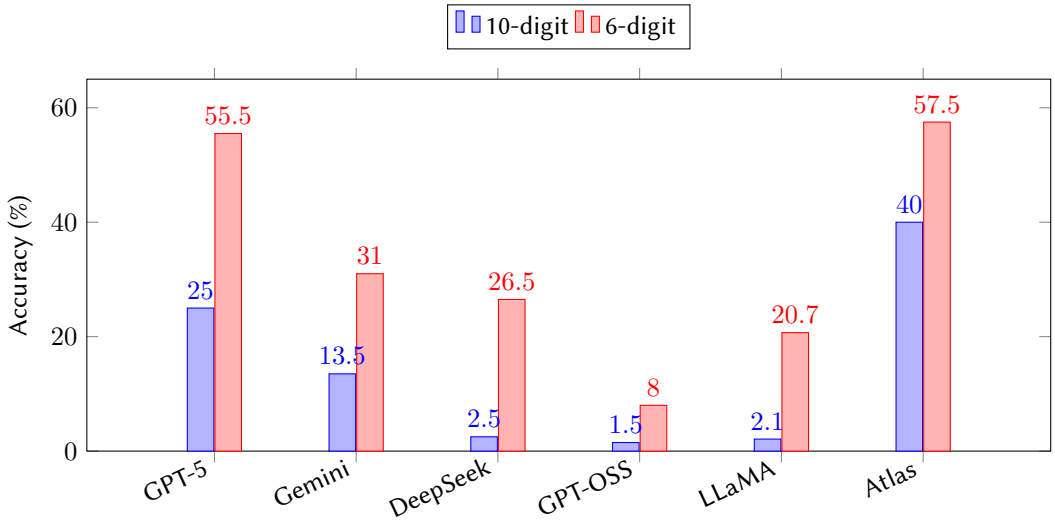


**Figure 2:** Comparison of model accuracy at 10- and 6-digit levels.

Interestingly, GPT-5-Thinking and Atlas are much closer at the 6-digit (globally harmonized) level than at the full 10-digit US level (55.5% vs. 57.5%). We view this as evidence that large proprietary models already capture a substantial amount of generic semantic knowledge about product categories and HS chapters, while the remaining performance gap at 10 digits is driven by U.S.-specific subheadings and edge cases that benefit disproportionately from domain-adapted fine-tuning on CROSS rulings. In other words, Atlas provides most of its value in the last few digits of the hierarchy and in its open-weight, cost-efficient deployability, rather than simply re-learning globally harmonized structure that frontier models already approximate.

## 4.2. Inference Cost

Cost per classification is crucial for scalability. Table 4 shows that open-source models, particularly Atlas, are an order of magnitude cheaper while maintaining state-of-the-art accuracy.

Cost estimates for proprietary models are based on public API pricing as of September 2025 for prompt and completion tokens at the context lengths used in our experiments (see footnote in Section 4). For open-weight models, including the base LLaMA-3.3-70B and Atlas, we report an approximate per-1,000 inference cost assuming on-premise deployment on A100-80GB GPUs with standard cloud pricing. While absolute numbers will vary with hardware, batch size, and provider, the relative ordering (frontier APIs being several times more expensive than open-weight deployment at similar throughput) is robust.

| Model | Cost per 1,000 inferences (USD) |
|---|---|
| GPT-5-Thinking | 3.30 |
| Gemini-2.5-Pro-Thinking | 5.50 |
| DeepSeek-R1 | 1.00 |
| GPT-OSS-120B | 0.90 |
| LLaMA-3.3-70B | 0.70 |
| **Atlas (fine-tuned)** | **0.70** |

**Table 4**
Estimated cost for 1,000 HTS inferences.

### 4.3. Discussion

Taken together, these results highlight a critical tradeoff: **Atlas** not only surpasses GPT-5-Thinking in accuracy (40% vs. 25% fully correct classifications), but also reduces inference cost by nearly **5× compared to GPT-5** and almost **8× compared to Gemini-2.5-Pro-Thinking**. Moreover, the strong performance on partially correct classifications demonstrates that Atlas generalizes beyond U.S.-specific tariffs to the globally harmonized 6-digit regime, reinforcing its utility for international trade applications. A qualitative comparison illustrating model reasoning differences is provided in Appendix C.

## 5. Summary and Future Directions

We introduced the first benchmark for Harmonized Tariff Schedule (HTS) code classification and presented Atlas, a fine-tuned LLaMA-3.3-70B model for global trade compliance. The study establishes HTS classification as a challenging new LLM benchmark, with three main takeaways:

- **Performance:** Atlas achieves 40% fully correct and 57.5% partially correct (6-digit) classifications, surpassing GPT-5-Thinking (+15 pts) and Gemini-2.5-Pro (+27.5 pts).
- **Efficiency:** Atlas is **5× cheaper than GPT-5** and **8× cheaper than Gemini**, while supporting secure self-hosted deployment.
- **Challenge:** Even the best model attains only 40% 10-digit accuracy, underscoring substantial headroom for progress.

Future work includes expanding coverage beyond semiconductors, distilling Atlas into smaller (8B–3B) variants for edge use, and applying reinforcement learning via rule-based rewards [16, 18] to improve reasoning and alignment.

We release Atlas as an open-weight model, allowing organizations to self-host the model under their own compliance and privacy constraints.

## Declaration on Generative AI

During the preparation of this work, the author(s) used large language models (e.g., OpenAI GPT-4 and GPT-5) for grammar checking, wording refinement, technical editing, and assistance in restructuring some paragraphs. No generative AI tools were used to create figures, tables, or experimental results. After using these tools, the author(s) reviewed, verified, and edited all generated content as needed and take full responsibility for the publication's final text and accuracy.

## References

[1] United States International Trade Commission, Harmonized tariff schedule (hts us), https://hts.usitc.gov/, 2025. Accessed: 2025-09-20.

[2] India temporarily suspends most postal services to us effective august 25 amid new customs order, Times of India (2025). URL: https://timesofindia.indiatimes.com/india/india-temporarily-suspends-most-postal-services-to-us-effective-august-25-amid-new-customs-order-know-what articleshow/123469918.cms.

[3] Dhl, german postal service suspend transport of parcels to us, Reuters (2025). URL: https://www.reuters.com/business/dhl-german-postal-service-suspend-transport-business-parcels-us-2025-08-22/.

[4] Countries suspend postal shipments to the us: full list, USA Today (2025). URL: https://www.usatoday.com/story/money/2025/08/28/countries-suspended-postal-shipments-to-us-list/85867109007/.

[5] P. Yuvraj, S. Devarakonda, Cross rulings hts dataset for tariffs, https://huggingface.co/datasets/flexifyai/cross_rulings_hts_dataset_for_tariffs, 2025.

[6] P. Yuvraj, S. Devarakonda, Atlas: Benchmarking and adapting llms for global trade via harmonized tariff code classification, https://huggingface.co/flexifyai/atlas-llama3.3-70b-hts-classification, 2025.

[7] U.S. Customs and Border Protection, Customs rulings online search system, https://rulings.cbp.gov/home, 2025. Accessed: 2025-09-20.

[8] Selenium Project, Selenium with python, https://www.selenium.dev/documentation/, 2025. Accessed: 2025-09-20.

[9] Google, Chromedriver, https://chromedriver.chromium.org/, 2025. Accessed: 2025-09-20.

[10] S. Pirogov, Webdriver manager for python, https://github.com/SergeyPirogov/webdriver_manager, 2025. Accessed: 2025-09-20.

[11] OpenAI, A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al., Gpt-4o system card, arXiv preprint arXiv:2410.21276 (2024). URL: https://arxiv.org/abs/2410.21276.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, arXiv preprint arXiv:2201.11903 (2023). URL: https://arxiv.org/abs/2201.11903.

[13] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhotia, L. Rantala-Yeary, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet,

V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan, I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, Z. Ma, The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[14] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in Neural Information Processing Systems 33 (2020) 1877–1901.

[15] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems 35 (2022) 27730–27744.

[16] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, D. Guo, Deepseekmath: Pushing the limits of mathematical reasoning in open language models, arXiv preprint arXiv:2402.03300 (2024). URL: https://arxiv.org/abs/2402.03300.

[17] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347 (2017). URL: https://arxiv.org/abs/1707.06347.

[18] R. Rafailov, A. Sharma, E. Mitchell, C. Zhang, C. D. Manning, C. Finn, S. Ermon, Direct preference

optimization: Your language model is secretly a reward model, Advances in Neural Information Processing Systems 36 (2023).

# A. Full Dataset Distribution

Table 5 lists the complete distribution of CBP rulings across all included Harmonized Tariff Schedule (HTS) chapters. This extended version complements the abbreviated table in Section 2.

**Table 5**
Full distribution of CBP rulings across HTS chapters.

| Chapter | HTS Codes | Rulings | Description |
|---|---|---|---|
| 27 | 3 | 7 | Mineral fuels, mineral oils and products of their distillation |
| 28 | 55 | 253 | Inorganic chemicals; organic or inorganic compounds of precious metals |
| 29 | 324 | 1199 | Organic chemicals |
| 32 | 16 | 44 | Tanning or dyeing extracts; dyes, pigments, tannins and derivatives |
| 35 | 23 | 306 | Albuminoidal substances; modified starches; glues; enzymes |
| 39 | 264 | 2781 | Plastics and articles thereof |
| 40 | 89 | 417 | Rubber and articles thereof |
| 42 | 28 | 223 | Articles of leather; saddlery, harness, travel goods, handbags |
| 49 | 1 | 12 | Printed books, newspapers, pictures, and other printed products |
| 61 | 163 | 3445 | Apparel and clothing accessories, knitted or crocheted |
| 70 | 17 | 36 | Glass and glassware |
| 72 | 10 | 15 | Iron and steel |
| 73 | 389 | 1749 | Articles of iron or steel |
| 74 | 52 | 119 | Copper and articles thereof |
| 76 | 27 | 80 | Aluminum and articles thereof |
| 82 | 152 | 1450 | Tools, implements, cutlery, spoons, and forks of base metal |
| 83 | 6 | 17 | Miscellaneous articles of base metal |
| 84 | 801 | 3566 | Nuclear reactors, boilers, machinery, and mechanical appliances |
| 85 | 293 | 1445 | Electrical machinery and equipment; sound recorders and reproducers |
| 87 | 17 | 60 | Vehicles (other than railway/tramway rolling stock) and parts |
| 90 | 256 | 1499 | Optical, photographic, measuring, precision, and medical instruments |
| 94 | 6 | 8 | Furniture; bedding, mattresses, cushions, and similar furnishings |
| **All** | 2992 | 18731 | All chapters combined |

## B. Prompt Template for Data Transformation

Each ruling was converted into a structured instruction–response pair to facilitate supervised fine-tuning of language models. The complete transformation prompt is shown below.

```
Given the following HTS ruling information:

HTS Code: {hts_code}
Ruling Number: {ruling_number}
Title: {title}
Date: {date}
URL: {url}
Summary: {summary}
Content: {content}

Please analyze this information and provide:

a) A concise product description representing the item being classified
b) A reasoning path justifying why the HTS US code is correct
c) The final HTS US code

Format your response as follows:

User: What is the HTS US Code for [product_description]?
Model:
HTS US Code -> [HTS US Code]
Reasoning -> [detailed_reasoning_path]
```

This design enforces explicit reasoning traces, aligning with recent advances in chain-of-thought modeling [12].

## C. Qualitative Example

Table 6 presents one representative example from the CROSS test set, showing how Atlas reasons through a real-world tariff classification case compared to other models. The excerpt highlights how Atlas distinguishes manufacturing stages and correctly identifies the applicable HTS code.

**Table 6**
Representative qualitative example from the CROSS dataset.

| | |
|---|---|
| **Product Description** | Silicon wafers, partially processed, used for semiconductor fabrication. |
| **Ground Truth HTS Code** | 3818.00.00.00 — Chemical elements doped for use in electronics. |
| **Atlas (Ours)** | *Reasoning:* Identifies item as doped wafer, not a finished semiconductor device; recognizes manufacturing stage. *Prediction:* 3818.00.00.00 (✓) Fully correct. |
| **GPT-5-Thinking** | *Reasoning:* Focuses on "semiconductor fabrication," misclassifies as complete integrated circuit. *Prediction:* 8542.31.00.00 (×) Incorrect; classifies as final chip. |
| **Gemini-2.5-Pro-Thinking** | *Reasoning:* Associates with silicon materials but ignores doping context. *Prediction:* 3824.99.99.99 (×) Incorrect; generic chemical compound. |