ATLAS: BENCHMARKING AND ADAPTING LLMS FOR GLOBAL TRADE VIA HARMONIZED TARIFF CODE CLASSIFICATION

Anonymous authors
Paper under double-blind review

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.

[1]

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 Commission (2025).

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 tim (2025); reu (2025); usa (2025), 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:

[1] 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`

- The first open-source benchmark for HTS classification Yuvraj & Devarakonda (2025a), derived from the U.S. Customs Rulings Online Search System (CROSS);
- 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 Yuvraj & Devarakonda (2025b) (LLaMA-3.3-70B), achieving 40% 10-digit and 57.5% 6-digit accuracy—substantially outperforming baselines—while being up to 8× 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.

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) Customs & Protection (2025). 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.

2.1 Collection and Scope

We built an automated agent Project (2025); Google (2025); Pirogov (2025) to scrape CROSS and align each ruling with its official 10-digit HTS code from Commission (2025). 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 OpenAI et al. (2024) 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 inAppendix B. This structure enforces reasoning-based prediction, aligning with chain-of-thought research Wei et al. (2023).

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 Yuvraj & Devarakonda (2025a).

Table 2: CROSS dataset splits.

Training	18,254
Validation	200
Test	200

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** Grattafiori et al. (2024). 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) Brown et al. (2020); Ouyang et al. (2022). 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 $x = (x_1, \dots, x_n)$ and target sequence $y = (y_1, \dots, y_m)$, the model with parameters θ 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 1×10^{-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 × 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.

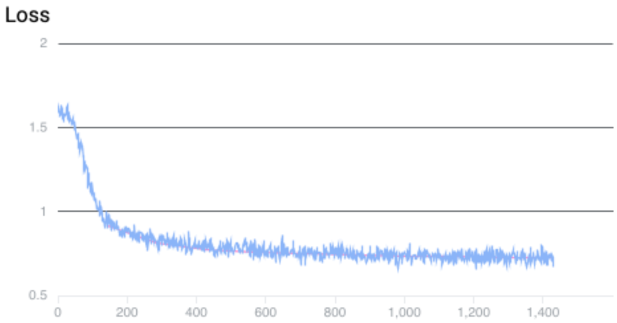

Figure 1: Training loss curve over 1,400 optimization steps. Rapid early improvement is followed by stable convergence.

3.3 Beyond Supervised Fine-Tuning: Reinforcement Learning

Although SFT provides a strong domain-adapted baseline, it is inherently limited by the distribution of observed CROSS rulings. Reinforcement learning (RL) offers a pathway to improve generalization beyond exact supervision.

A lightweight and cost-effective starting point is 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 Shao et al. (2024) or related policy-gradient methods such as PPO Schulman et al. (2017). 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.

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 trade-off 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) Rafailov et al. (2023) 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).

4.1 Accuracy at 10 and 6 Digits

Table 3 shows fully correct classifications. GPT-5-Thinking[2] 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 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.

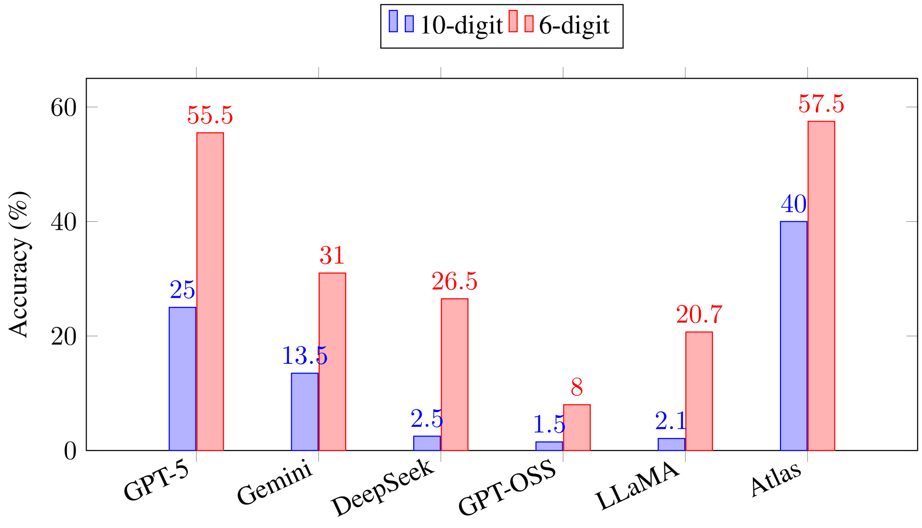

Figure 2: Comparison of model accuracy at 10- and 6-digit levels.

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.

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

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 Shao et al. (2024); Rafailov et al. (2023) to improve reasoning and alignment.

We release Atlas Yuvraj & Devarakonda (2025b) and the dataset Yuvraj & Devarakonda (2025a) to encourage research on domain-specialized, trustworthy LLMs for global trade.

References

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/`.

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 articleshow/123469918.cms`.

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/`.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

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

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

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide

Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Sergey Pirogov. Webdriver manager for python. https://github.com/SergeyPirogov/webdriver_manager, 2025. Accessed: 2025-09-20.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher Zhang, Christopher D Manning, Chelsea Finn, and Stefano Ermon. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

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

Pritish Yuvraj and Siva Devarakonda. Cross rulings hts dataset for tariffs. https://huggingface.co/datasets/flexifyai/cross_rulings_hts_dataset_for_tariffs, 2025a.

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

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 Wei et al. (2023).

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.