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

Pritish Yuvraj
Flexify.AI
pritish@flexify.ai

Siva Devarakonda*
Flexify.AI
siva@flexify.ai

ABSTRACT

Accurate classification of products under the Harmonized Tariff Schedule (HTS) is a critical bottleneck in global trade, yet it has received little attention from the machine learning community. Misclassification can halt shipments entirely, with major postal operators suspending deliveries to the U.S. due to incomplete customs documentation.

We introduce the first benchmark for HTS code classification, derived from the U.S. Customs Rulings Online Search System (CROSS). Evaluating leading LLMs, we find that our fine-tuned ATLAS model (LLaMA-3.3-70B) achieves **40% fully correct 10-digit classifications** and **57.5% correct 6-digit classifications**—improvements of **+15 points over GPT-5-Thinking** and **+27.5 points over Gemini-2.5-Pro-Thinking**.

Beyond accuracy, ATLAS is **5× cheaper than GPT-5-thinking** and **8× cheaper than Gemini-2.5-Pro-Thinking**, and can be self-hosted to guarantee data privacy—an essential requirement in high stake industries like Automotives ,Industrials, Semiconductors etc. for trade and compliance workflows. While ATLAS sets a strong baseline, the benchmark remains highly challenging, with only 40% 10-digit accuracy.

By releasing both dataset and model, we aim to position HTS classification as a new community benchmark task. We invite future work in retrieval, reasoning and alignment to advance progress on this high-impact global trade problem.

[1]

1 INTRODUCTION

Every product imported into the global market must be assigned a Harmonized Tariff Schedule (HTS) code. These codes, standardized by the World Customs Organization (WCO), are ten digits long. The first six digits are harmonized across all participating countries, while the last four digits are country-specific extensions. Correctly identifying the first six digits enables global interoperability, while the full ten-digit code is required for compliance with U.S. customs.

The HTS is deeply hierarchical: 22 sections are divided into 99 chapters, which expand into thousands of subheadings. Chapters 1–97 correspond to stable product categories (such as chemicals, machinery, and textiles), whereas Chapters 98–99 capture temporary and special provisions that change frequently. This structure makes tariff classification a natural hierarchical machine learning task, where six-digit accuracy captures worldwide consistency, and ten-digit accuracy reflects the U.S.-specific extension.

*Website: https://tariffpro.flexify.ai/

[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
3. Atlas LLM Model Demo: https://flexifyai-atlas-llama3-3-70b-hts-demo.hf.space/?logs=container&__theme=system&deep_link=FFJuTJsv_fM

Despite its centrality, classification remains a major bottleneck. Recent trade policy changes, for example, the modifications to the *de minimis* exemption, require that any imported good valued above $100 must be assigned a valid HTS code. The HTS itself spans more than 17,000 pages of PDF documents, making manual assignment infeasible at scale. The consequences are global: in 2025, several major postal operators suspended parcel delivery to the United States, citing the inability to assign correct HTS codes and complete customs documentation tim (2025), reu (2025), usa (2025). More than thirty countries were affected, highlighting the fragility of global trade flows when classification is not available at scale.

Large Language Models (LLMs) offer a scalable alternative. Their capacity for semantic reasoning and structured classification makes them natural candidates for HTS code classification, where fine-grained distinctions must be captured. Moreover, since the first six digits are harmonized worldwide, advances in HTS classification can simultaneously benefit global trade systems, while the U.S.-specific digits directly address compliance in the American market.

1.1 CONTRIBUTIONS

Our key contributions are:

- We release the first open-source benchmark for HTS classification Yuvraj & Devarakonda (2025a), constructed from the U.S. Customs Rulings Online Search System (CROSS), including training, validation, and test splits.
- We benchmark leading proprietary and open-source models, including GPT-5-Thinking OpenAI (2025a), Gemini-2.5-Pro-Thinking DeepMind (2025), LLaMA-3.3-70B Grattafiori et al. (2024), DeepSeek-R1 (05/28) DeepSeek-AI et al. (2025), and GPT-OSS-120B OpenAI (2025b).
- We fine-tuned LLaMA-3.3-70B with supervised fine-tuning (SFT) to create the specialized model ATLAS, which we open source Yuvraj & Devarakonda (2025b). ATLAS achieves **40% accuracy at the 10-digit level** and **57.5% at the 6-digit level**, substantially outperforming GPT-5-Thinking (25%) and Gemini-2.5-Pro-Thinking (13.5%).
- Beyond accuracy, ATLAS is significantly more cost-efficient—up to **8× cheaper per inference**—and supports privacy-preserving deployment through self-hosting, ensuring that sensitive trade data never leaves secure environments.

Together, these contributions establish tariff code classification as a new frontier for LLM evaluation, situated at the heart of compliance for Global Commerce and Trade.

1.2 PAPER ROADMAP

The remainder of this paper is organized as follows. Section 2 describes the construction of our CROSS-based dataset and its transformation into a machine-learning–ready format. Section 3 details the fine-tuning procedure for ATLAS. Section 4 presents evaluation results across multiple proprietary and open-source LLMs, analyzing both accuracy and cost efficiency. Finally, we conclude with a summary of key findings and future research directions in Section 5.

2 DATASETS

A central contribution of this work is the construction of 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 decisions issued by U.S. Customs and Border Protection (CBP), in which importers or brokers sought clarification on the correct HTS code for specific products. These rulings are authoritative, high-value examples of tariff classification in practice, but are lengthy, unstructured, and scattered across thousands of HTML pages—making them inaccessible for machine learning research.

2.1 DATA COLLECTION

We developed an automated browser agent Project (2025); Google (2025); Pirogov (2025) to systematically scrape CROSS. Each ruling was matched to a 10-digit HTS code obtained from the official HTS U.S. website Commission (2025). After filtering and cleaning, the final dataset spans 18,731 rulings covering 2,992 unique HTS codes across a broad range of product categories.

Not every HTS code appears in CROSS, since only disputed or clarified codes are documented. However, the presence of a code in CROSS is itself informative: frequent rulings signal categories that are high-demand or ambiguous in practice, while absent codes suggest stable or rarely used classifications.

2.2 DATA TRANSFORMATION INTO LLM-TRAINABLE FORMAT

Raw CROSS rulings are official letters—legalistic, verbose, and inconsistent in structure. To make them suitable for supervised learning, we transformed each ruling into a structured prompt–response format using GPT-4o-mini OpenAI et al. (2024). This lightweight model was cost-effective and sufficient for information extraction.

Prompt template. Each ruling was converted into the following instruction format:

```
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) Create a product description that the user was initially getting the HTS US code f
b) Create a reasoning path justifying why the HTS US code is correct
c) Return the 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 forces models to both predict the code and provide a reasoning path, aligning with recent work on chain-of-thought reasoning Wei et al. (2023).

2.3 DATASET SPLITS

From the 18,731 processed rulings, we randomly sampled 200 examples for validation and 200 for final testing, holding them out strictly from training. The remaining 18,254 rulings form the training set. This ensures a clean separation between model development and final evaluation. We have uploaded the dataset to hugging-face Yuvraj & Devarakonda (2025a).

2.4 DISCUSSION

This dataset poses unique challenges: (1) rulings are lengthy and often hinge on subtle distinctions (e.g., partially fabricated vs. fully fabricated semiconductor wafers); (2) correctness has a hierarchical structure (6-digit vs. 10-digit); and (3) errors carry real-world consequences for trade and com-

Table 1: Distribution of the CROSS dataset into training, validation, and test splits.

Split	Number of Rulings
Training	18,254
Validation	200
Test	200

pliance. By releasing this benchmark, we aim to establish tariff classification as a novel, high-impact evaluation task for LLMs, complementing existing benchmarks in reasoning, code generation, and multilingual understanding.

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 $x = (x_1, \ldots, x_n)$ and target sequence $y = (y_1, \ldots, y_m)$, 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 $\times$ 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.

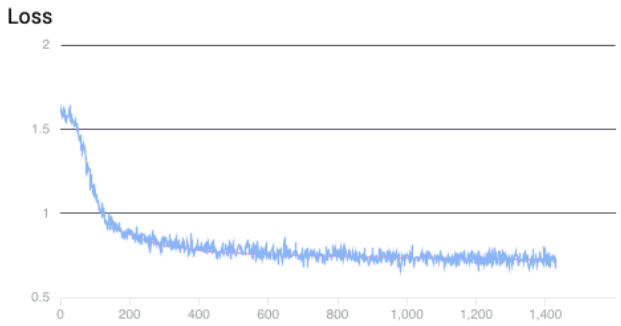

Figure 1: Training loss curve over 1,400 optimization steps. Rapid early improvement is followed by stable convergence.

3.3 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.

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. The task requires classifying the correct 10-digit HTS US code for each product description. Since tariff classification is inherently hierarchical, we report three complementary evaluation metrics:

- **Fully correct classification:** all 10 digits match exactly. A fully correct classification means that the end-to-end classification pipeline produces an operationally valid HTS US code, enabling the product to clear customs.
- **Partially correct classification:** the first 6 digits (harmonized across all WTO members) match. This reflects whether the model generalizes to the globally standardized portion of the code, making it directly relevant for cross-border deployments outside the U.S.
- **Average digit-level accuracy:** mean number of correctly predicted digits (0–10), capturing fine-grained improvements even when full correctness is not achieved.

4.1 Fully Correct Classifications

Table 2 reports the number and percentage of classifications that exactly match the 10-digit HTS US code. General-purpose LLMs such as GPT-5-Thinking achieve moderate success (25%), whereas open-source baselines without domain adaptation perform poorly ($\leq$ 3%). Our fine-tuned model,

Atlas, based on LLaMA-3.3-70B, achieves the best results with 40% fully correct classifications—meaning nearly half of all test products are classified into a customs-ready code.

Model	Fully Correct (N)	Accuracy (%)
GPT-5-Thinking	50	25.0%
Gemini-2.5-Pro-Thinking	27	13.5%
DeepSeek-R1 (05/28)	5	2.5%
GPT-OSS-120B	3	1.5%
LLaMA-3.3-70B	3	2.1%
Atlas (Fine-tuned LLaMA-3.3-70B)	**80**	**40.0%**

Table 2: Fully correct HTS US code classifications (10-digit match) on the 200-sample test set.

4.2 Partially Correct Classification

Table 3 evaluates classifications at the 6-digit level, which is harmonized globally and thus forms the basis for international tariff schedules. Here, GPT-5-Thinking reaches 55.5% accuracy, while **Atlas** achieves 57.5%. This shows that our domain-specific fine-tuning not only improves U.S.-specific classification but also transfers to the globally harmonized layer, demonstrating potential for adoption in worldwide trade contexts.

Model	Partially Correct (N)	Accuracy (%)
GPT-5-Thinking	111	55.5%
Gemini-2.5-Pro-Thinking	62	31.0%
DeepSeek-R1 (05/28)	53	26.5%
GPT-OSS-120B	16	8.0%
LLaMA-3.3-70B	29	20.7%
Atlas (Fine-tuned LLaMA-3.3-70B)	**115**	**57.5%**

Table 3: Partially correct HTS code classifications (6-digit harmonized match).

4.3 Average Digit-Level Accuracy

Finally, we report the average number of digits correctly predicted per code in Table 4. While general-purpose models hover around 3–5 digits correct, **Atlas** achieves 6.3 digits correct on average. This demonstrates that supervised fine-tuning on CROSS rulings improves fine-grained reasoning over tariff codes, even when full correctness is not reached.

Model	Avg. Digits Correct (out of 10)
GPT-5-Thinking	5.61
Gemini-2.5-Pro-Thinking	2.92
DeepSeek-R1 (05/28)	3.24
GPT-OSS-120B	2.58
LLaMA-3.3-70B	3.31
Atlas (Fine-tuned LLaMA-3.3-70B)	**6.30**

Table 4: Average number of correctly predicted digits per HTS code.

4.4 Visual Comparison

To complement the tables, Figure 2 summarizes model performance across both evaluation levels. Atlas's advantage is especially pronounced at the 10-digit U.S.-specific classification task.

4.5 Cost Efficiency of Inference

Beyond accuracy, cost per inference is a critical factor for practical deployment of tariff classification systems. Closed-source API models such as GPT-5-Thinking and Gemini-2.5-Pro-Thinking incur

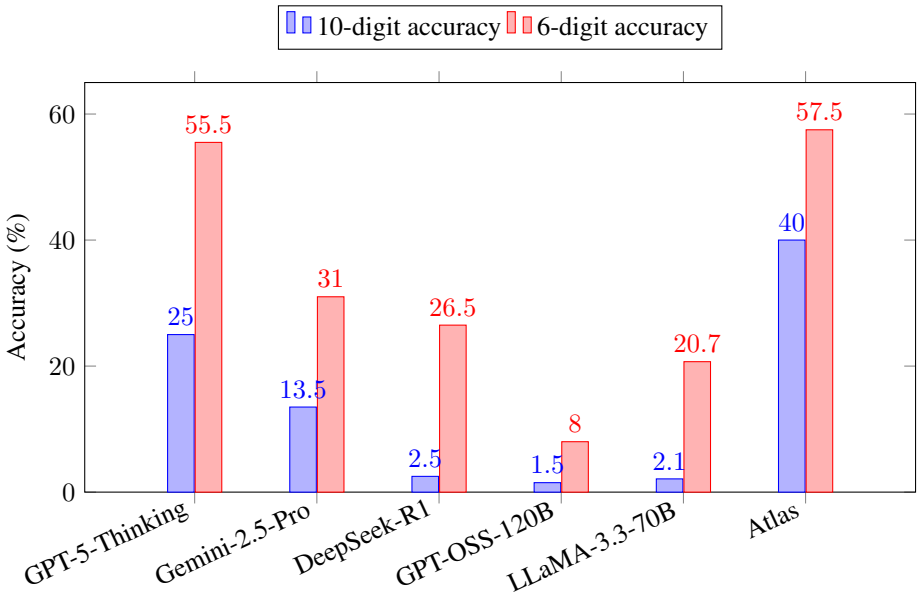

Figure 2: Visual comparison of model performance on HTS classification. Atlas leads at both evaluation levels, with a marked margin at the 10-digit (U.S.-specific) level.

substantial per-query costs, particularly when scaled to thousands of classifications, whereas fine-tuned open-source models can be hosted locally or on cost-effective GPU clusters at a fraction of the price.

Table 5 compares the cost of classifying 1,000 product descriptions into 10-digit HTS codes. We assume a standard context length ($\sim$1k input tokens, $\sim$200 output tokens) and use publicly available API pricing at the time of writing. For open-source models (LLaMA-3.3-70B and Atlas), costs are estimated from on-demand A100 GPU cloud pricing.

Model	**Cost for 1,000 HTS Inferences (USD)**
GPT-5-Thinking	$\sim$ \$3.30
Gemini-2.5-Pro-Thinking	$\sim$ \$5.50
DeepSeek-R1 (05/28)	$\sim$ \$1.00
GPT-OSS-120B	$\sim$ \$0.90 (estimated compute)
LLaMA-3.3-70B	$\sim$ \$0.70 (self-hosted)
Atlas (Fine-tuned LLaMA-3.3-70B)	$\sim$ \$0.70 **(self-hosted)**

Table 5: Estimated cost of classifying 1,000 products into 10-digit HTS codes. Closed-source models use API pricing; open-source models assume on-demand A100 GPU hosting.

4.6 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$\times$ compared to GPT-5** and almost **8$\times$ 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.

5 Summary and Future Directions

This work introduced the first real world benchmark for Trade policy reasoning based on Harmonized Tariff Schedule (HTS) code classification and presented Atlas, a fine-tuned LLaMA-3.3-70B

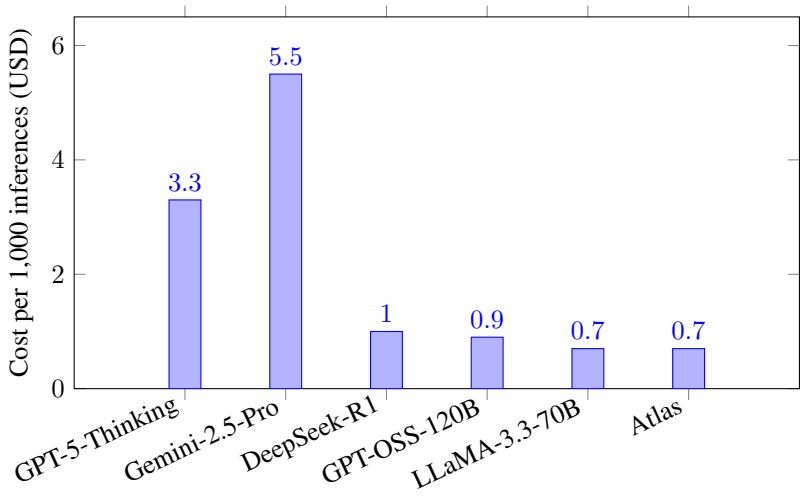

Figure 3: Estimated cost per 1,000 HTS inferences. Atlas (self-hosted) is substantially cheaper than proprietary APIs.

model adapted to this high-stakes domain. Our study establishes tariff classification as a challenging new frontier for LLM evaluation, with three central takeaways:

- **State-of-the-art performance:** ATLAS achieves 40% fully correct classifications at the 10-digit level and 57.5% at the 6-digit level, outperforming GPT-5-Thinking (+15 points) and Gemini-2.5-Pro-Thinking (+27.5 points).
- **Cost and deployment efficiency:** ATLAS is nearly **5× cheaper than GPT-5** and **8× cheaper than Gemini**, while enabling self-hosted deployment for sensitive trade and supply-chain applications.
- **Open benchmark challenge:** Despite these gains, best 10-digit accuracy remains only 40%, underscoring the need for advances in reasoning, retrieval, and alignment methods.

Looking forward, we see three promising directions: (1) expanding the dataset to include a broader range of rulings beyond the current subset, (2) distilling ATLAS into smaller variants (e.g., 8B or 3B) for efficient deployment in resource-constrained settings, and (3) exploring enhanced reasoning techniques and retrieval-augmented methods to improve classification accuracy.

We release ATLAS Yuvraj & Devarakonda (2025b) and the benchmark splits on Hugging Face Yuvraj & Devarakonda (2025a) to support reproducibility. By framing HTS classification as a benchmark task, we aim to catalyze progress on domain-specialized LLMs—directly tied to the resilience of global trade and supply chains.

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 DeepMind. Gemini 2.5 pro thinking. https://deepmind.google/models/gemini/pro/, 2025. Accessed: 2025-09-20.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

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. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, 2025a. Accessed: 2025-09-20.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025b. URL https://arxiv.org/abs/2508.10925.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, 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.

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.