LLM-Based Product Recommendation with Prospect Theoretic Self Alignment Strategy

Manying Zhang
Institut National des Langues
et Civilisations Orientales
manying.zhang@inalco.fr

Zehua Cheng
University of Oxford
zehua.cheng@cs.ox.ac.uk

Damien Nouvel
Institut National des Langues
et Civilisations Orientales
damien.nouvel@inalco.fr

Abstract

Accurate and personalized product recommendation is central to user satisfaction in e-commerce. However, a persistent language gap often exists between user queries and product titles or descriptions. While traditional user behavior-based recommenders and LLM-based Retrieval-Augmented Generation systems typically optimize for maximum likelihood objectives, they may struggle to bridge this gap or capture users' true intent. In this paper, we propose a strategy based on Prospect Theoretic Self-Alignment, that reframes LLM-based recommendations as a utility-driven process. Given a user query and a set of candidate products, our model acts as a seller who anticipates latent user needs and generates product descriptions tailored to the user's perspective. Simultaneously, it simulates user decision-making utility to assess whether the generated content would lead to a purchase. This self-alignment is achieved through a training strategy grounded in Kahneman & Tversky's prospect theory, ensuring that recommendations are optimized for perceived user value rather than likelihood alone. Experiments on real-world product data demonstrate substantial improvements in intent alignment and recommendation quality, validating the effectiveness of our approach in producing personalized and decision-aware recommendations.

1 Introduction

In e-commerce, personalized product recommendations have become essential to enhancing user experience. Traditional systems relied on user behavior and collaborative filtering to suggest items, while conversational recommendation systems enabled direct expression of user needs. More recently, large language models (LLMs) have expanded recommendation capabilities by simplifying the semantic understanding of product queries, even without user history. A common practice is the Retriever-Augmented Generation approach where user queries are first made into embeddings to match the most similar product representations, then recommendations are based on these selected products. However, despite making large scale product filtering possible and making interactions more active, LLM systems can inadvertently mislead consumers through information echo chambers or inaccurate recommendations due to the language gap between user queries and product contents. One intuitive reason for this phenomenon is rooted in how LLMs are trained: with Maximum Likelihood Estimation (MLE) as the primary objective, models often learn to mimic superficial patterns and reproduce ”false personalization” inherent in the training data.

In this paper, we introduce a novel approach to address these issues by leveraging Kahneman & Tversky's Prospect Theory from psychology. Prospect Theory provides a framework that better reflects real human behavior compared to the rational agent hypothesis, incorporating concepts such as loss aversion and preference reversal (Tversky and Kahneman, 1992). Inspired by this, we apply Kahneman & Tversky's prospect theoretic optimization (Ethayarajh et al., 2024) to train our LLM-based model ”Oracle” with the target of maximizing the purchase utility of its generated product content, rather than optimizing for log-likelihood. Our self-alignment mechanism is grounded in a dual-role modeling: Given a user query and related products, our model first anticipates latent user needs and generates unbiased product descriptions. It then simulates user decisions—such as add to cart, abandon or purchases—on the generated content to evaluate its effectiveness. These simulated decisions act as implicit utility signals, which guide the model to iteratively refine its generation behavior. This self-supervised, utility-driven loop enables scalable and principled alignment between

Proceedings of Recent Advances in Natural Language Processing,pages 1430–1436
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-165

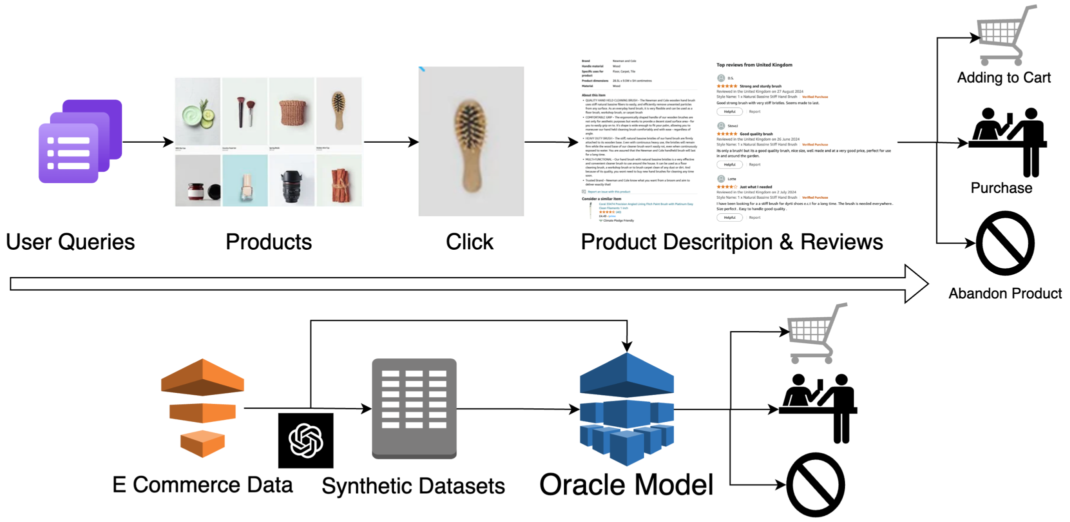

Figure 1: User behavior (upper half) and the training logic and structure of the Oracle model. The final user action choices are: adding to cart, purchasing, or abandoning the product. We fine-tuned the Qwen2-72B model to built our Oracle model based on e-commerce data to predict the click-through rate of specific products after a given query. We assume that users read all titles, descriptions, and reviews for each product. We also use GPT-4o-2024-05-13 (OpenAI, 2023) to generate synthetic data to improve the diversity of dataset.

model outputs and user value. Our contribution can be summarized as below:

- We proposed a novel LLM-based generation and simulator framework to align product recommendations with rational decision-making principles and released our code.[1].
- We integrate Kahneman and Tversky's Prospect Theory into the optimisation of our model, enabling it to generate unbiased product content that not only captures latent consumer needs but also bridges the gap between aggressive marketing tactics and rational decision-making for both consumers and sellers.

2 Related Work

2.1 User simulator and shopping assistant by LLMs

Recently LLMs have been employed to simulate user behaviors, aiding in the development and evaluation of e-commerce recommendations. For instance, LLMs are used to model diverse user profiles, capturing varying needs and personalities to support conversational sales agents in strategic decision-making (Kim et al., 2025). Similarly, (Kasuga and Yonetani, 2024) leverage LLM embeddings to predict user behavior transitions, facilitating the assessment of web-marketing campaigns without extensive online testing. The user behavior alignment challenge proposed by KDD (ama, 2024) introduces the ShopBench dataset (Jin et al., 2024b), comprising approximately 20,000 questions across 57 different tasks, including multiple-choice, retrieval, generation, and ranking questions to evaluate models' abilities to understand and align with implicit and heterogeneous user behaviors in online shopping environments. The winning team fine-tuned a LLM using extensive synthetic data generation and task augmentation strategies, achieving top performance in the challenge (Deotte et al., 2024).

On the other hand, LLMs are increasingly being integrated into shopping assistants to provide personalized and interactive e-commerce experiences. For instance, with the construction of instruction-version e-commerce datasets, researchers fine-tuned LLMs to enhance their capabilities in understanding product information and user reviews, thereby providing more accurate and context-aware shopping assistance (Peng et al., 2024; Li et al., 2023).

2.2 Kahneman-Tversky Optimization

Prospect Theory (Tversky and Kahneman, 1992) provides a model of human decision-making, focusing on how people perceive gains and losses differently through value functions, rather than absolute outcomes. The Kahneman-Tversky Optimization

[1] https://anonymous.4open.science/r/shopping-kto-5B4B

(KTO) model (Ethayarajh et al., 2024) builds on this by requiring only a binary feedback signal to guide the training objective. This feedback helps the model adjust to maximize overall utility based on perceived outcomes. In the e-commerce environment, we propose that large models should optimize the value functions of both buyers and sellers simultaneously, ensuring a fair balance of interests on both sides.

The simplicity and effectiveness of this thumb-up or thumb-down approach have inspired further exploration, as seen in Binary Classifier Optimization (Jung et al., 2024). This method builds on KTO's principles, explaining its effectiveness and achieving similar alignment results by optimizing a binary classifier. This approach offers a practical and scalable way to align LLMs with human preferences, demonstrating the broader applicability of KTO-inspired methodologies.

3 Dataset Construction

We utilized multiple data sources and the samples were transformed into instruction-style prompts.

- Amazon-M2 (Jin et al., 2024a): A comprehensive dataset of Amazon user sessions, containing historical clicks and the current click, enriched with detailed product metadata.
- Amazon Reviews 2023 (Hou et al., 2024): over 500 million Amazon reviews and purchase verifications of products spanning 33 categories;
- ESCI-data (Reddy et al., 2022): shopping query and its relevant results pairs, supplemented with ESCI relevance judgments (Exact, Substitute, Complement, Irrelevant) to assess product relevance.
- ECInstruct (Peng et al., 2024): 116,528 samples derived from 10 widely performed e-commerce tasks across four categories;
- MMLU (Hendrycks et al., 2020): massive multitask dataset consisting of over 100k multiple-choice questions and their answers;
- Alpaca-Cleaned (Taori et al., 2023): a cleaned version of the original Alpaca Dataset released by Stanford.

In addition, we found partial overlap among the products in the Amazon e-commerce datasets. We mapped features such as ”historical clicks” from Amazon-M2 and ”queries” from ESCI to represent the ”user query”, and used features like ”verified purchase” from Amazon Reviews 2023 and ”relevance judgment” from ESCI as labels for ”user action”. We then constructed a joint table dataset based on the common ”current product” column, resulting in total 2.6 million rows (Table 1). This joint dataset served as the training data for our user behavior alignment model, Oracle.

Table 1: User Action Dataset Split Overview

Split	Training	Dev	Test	Total
Number of Samples	2,080,000	260,000	260,000	2,600,000
Proportion	90%	10%	10%	100%

4 Oracle Model Construction

We present the overall pipeline for constructing the Oracle model in Figure 1. The Oracle model is trained to perform user behavior alignment on our large-scale e-commerce dataset (illustrated in the lower part of the figure). The objective is to predict user actions (purchase, add to cart, abandon) based on user queries (including natural language queries and historical clicks), the current product, as well as its descriptions and reviews, thereby simulating the decision-making process of a user (upper part).

The approach began with the collection and augmentation of diverse data, as described in Section 3. We processed the Amazon-M2 and Amazon Reviews 2023 datasets to extract and standardize user behavior signals such as product preferences, clickstreams, and implicit feedback. The ECInstruct and ESCI-data datasets provided structured e-commerce tasks, enabling the model to learn product categorization and relevance prediction. Additionally, the MMLU and Alpaca-Cleaned datasets contributed broader coverage for general reasoning and text generation.

Our joint dataset described in Section 3 was used to train the model's decision-making ability in shopping contexts. To enhance diversity, we also used GPT-4o-2024-05-13 (OpenAI, 2023) to generate synthetic data that covers under-represented user behaviors and product categories. These synthetic examples increased data diversity, helping the model generalize to a wider range of user interactions.

Details of the experimental setup are presented in Appendix **??**, and the instruction prompt design

Model	F1 Score w/o Synthetic data	w/ Synthetic data
Qwen1.5-14B	0.2003	0.2111
LlaMa3-70B-AWQ	0.6406	0.6771
ChatGPT-4o	0.8131	-
Smaug-72B-ZS	0.6564	0.6684
Smaug-72B-FT	0.6717	0.6907
Qwen2-72B-ZS	0.7783	0.7881
Qwen2-72B-FT	0.8113	**0.8235**

Table 2: Comparison different LLMs as zero-shot (ZS) with fine-tuned (FT) on the user behavior alignment task.

is elaborated in Appendix **??**.

This comprehensive data preparation and fine-tuning pipeline enabled the model to achieve high performance across various user behavior alignment tasks. Subsequently, we evaluated the shopping decision-making performance of the following models: Qwen1.5-14B (Qwen, 2023) (zero-shot), LLaMA3-70B-AWQ[2], Smaug-72B (Pal et al., 2024), Qwen2-72B-Instruct (Yang et al., 2024), and ChatGPT-4o-2024-05-13 (OpenAI, 2023). The results are presented in Table 2. Qwen2-72B-FT was selected as the Oracle model due to its superior performance. With the inclusion of synthetic data, Qwen2-72B-FT even outperformed ChatGPT-4o, achieving a confidence rate of 82% in user action prediction accuracy.

5 Prospect Theoretic Self-Alignment in e-Commerce

In the previous Section 4, we introduced our user simulator Oracle, which models user preferences over product content. We now focus on the core learning strategy, Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024), behind our self-alignment framework, which guides the content generator to produce product descriptions that maximize simulated user utility in our e-commercial recommendation setting.

5.1 Base Value Function in KTO

In the KTO framework, we aim to tune the model's parameters θ to maximize a utility function that incorporates human-like biases. A value function $v\left(z;\lambda,\alpha,z_0\right)$ maps an outcome z, relative to a reference point z_0 is formulated as:

$$v\left(z;\lambda,\alpha,z_0\right)=\begin{cases}\left(z-z_0\right)^{\alpha} & \text{if } z\geq z_0\\ -\lambda\left(z_0-z\right)^{\alpha} & \text{if } z<z_0\end{cases} \tag{1}$$

In this value function, z represents the actual outcome or gain/loss experienced by an individual, while z_0 denotes the initial expectation or reference point against which the individual evaluates gains or losses. If $z > z_0$, it is perceived as a gain; otherwise, it is perceived as a loss. This reference point is crucial, as it determines whether the individual perceives a situation as favorable or unfavorable. This function models the psychological process of decision-making under uncertainty, particularly in situations where individuals are faced with gains and losses. The two parameters α and λ play a crucial role in shaping the form and behavior of this value function.

- α: This parameter controls the curvature of the value function and is directly related to risk aversion. A lower value of α indicates that the individual is more sensitive to small variations in gains or losses. Conversely, a higher value of α implies a more linear behavior, reflecting less sensitivity to small variations, which can be interpreted as more risk-tolerant behavior.
- λ: This parameter determines loss sensitivity, or loss aversion. A value of $\lambda > 1$ indicates that individuals assign greater weight to losses than to gains of the same magnitude. In other words, losses are perceived as more significant than equivalent gains, reflecting the idea that people tend to avoid losses more strongly than they seek gains.

In our model, α and λ are adjusted to realistically reflect these human behaviors. For example, in consumer behavior studies, individuals are often found to exhibit strong loss aversion, which corresponds to λ values significantly greater than 1, while α may vary depending on the specific decision context.

The median values of the hyperparameters in our study are $\alpha = 0.88$ and $\lambda = 2.25$, which are representative values observed across individuals. These parameters directly influence how our model simulates decision-making processes and consumer preferences when faced with recommended products.

[2] https://huggingface.co/TechxGenus/Meta-Llama-3-70B-AWQ

Category	#Item	Base CVR	Qwen-ZS CVR	Qwen-FT CVR	ChatGPT-4o CVR	Qwen-KTSA CVR
All Beauty	112.6K	2.11	*2.08*	*2.08*	2.13	2.17
Appliances	94.3K	1.13	1.13	1.13	*1.10*	**1.30**
Baby Products	217.7K	2.71	*2.66*	*2.53*	*2.16*	**2.78**
Beauty & Care	1.0M	2.11	*1.90*	*1.91*	*2.01*	**2.13**
Books	4.4M	1.93	*1.90*	*1.91*	*1.90*	**1.96**
Apparence	7.2M	1.95	*1.90*	*1.93*	**2.11**	2.05
Pet Supplies	492.7K	2.14	*1.93*	*2.10*	**2.19**	**2.19**
Toys and Games	890.7K	2.85	*2.55*	*2.53*	*2.51*	*2.82*

Table 3: The average conversion rate (CVR) in E-Commerce with different setup of LLM for different categories. We followed the categories defined in Amazon Reviews 2023. All CVR is measured in percentage. The base conversion rate is measured and aggregated by the Oracle model (Qwen-2 72B-FT). Numbers marked in *italic* is lower then base CVR. The best CVR is marked in **bold** font.

5.2 Adapting KTO to E-Commerce

Since all existing content in e-Commerce is created by the seller, therefore all existing fine-tuning strategies without explicit regularization are considered as seller prospective. Therefore, the value function for seller is defined as $v(\cdot) = \cdot$, which is directly tuning model parameter based on the content.

For buyer prospective, the z_0 represents their initial expectations before engaging with a product. We defined the z_0 under our framework based on two aspects:

1. Average rating (R_{avg}): the numerical star ratings provided by previous buyers (normalized between $[0, 1]$);

2. Key features highlighted in reviews (X_{key}). Based on the reviews, we ask the LLM to generate the top-3 features and top-3 concerns (6 key features in total) in explicit natural language as product descriptions.

Hence, our z_0 could be defined as:

$$\max\left(R_{\text{avg}}, \frac{1}{6}\sum_{x_i \in X_{\text{key}}} \mathbb{E}[KL(\pi_{\theta'}(y|x)||\pi_\theta(y|x_i))]\right) \tag{2}$$

where $KL(\cdot||\cdot)$ is the Kullback–Leibler (KL) divergence. The $\pi_\theta(y|x)$ denotes the conditional probability distribution over possible outputs y given an input x, parameterized by θ. Here, y can be interpreted as the product descriptions generated by the model. The $\pi_{\text{ref}}(y|x)$ refers to the output distribution of the reference model, which in this case is the original pretrained model. The KL divergence measures the difference between our model and the reference model, reflecting the extent to which the generated descriptions diverges from the original product descriptions.

A large KL divergence indicates a significant deviation from the original descriptions. If the user proceeds to purchase the product despite such a divergence, this can be interpreted as the user ”appreciating” the change — suggesting that the new description better aligns with their expectations. In this case, the generated description and the product are considered “desirable”. Conversely, if the user does not purchase the product after a substantial change in the description, it implies that the newly generated description failed to meet their expectations and is thus “undesirable”. When users positively respond to a significantly different description, it may reveal that the original descriptions was biased or insufficient. In such cases, the model is encouraged to generate more objective and accurate descriptions. On the other hand, if users reject divergent contents, it suggests that the original content was already adequate, and the model should be constrained from introducing excessive deviation.

Therefore, in KTO's framework, we propose to reformulate Equation 1 as the following:

$$v(x,y) = \begin{cases} \lambda_D \sigma\left(\beta\left(r_\theta(x,y) - z_0\right)\right) \text{ if } y \sim y_{\text{desire}} \mid x \\ \lambda_U \sigma\left(\beta\left(z_0 - r_\theta(x,y)\right)\right) \text{ if } y \sim y_{\text{undire}} \mid x \end{cases} \tag{3}$$

where: $r_\theta(x, y) = \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$, which represents the difference between the two models in terms of the probability of generating a given description y for the same product, and is used to measure the divergence between the generated and original descriptions. λ_D and λ_U are hyperparameters for the ”desirable” and ”undesirable” losses, respectively. β is a number between 0 and 1 used for regularization. A sigmoid function $\sigma(\cdot)$ ensures that the value function remains bounded and behaves smoothly.

Finally, the overall value function is:

$$v(x, y) = v_{\text{buyer}}(x, y) + \gamma r_\theta(x, y) \tag{4}$$

where γ is a regularization factor. In this paper, we set all $\gamma = 1$.

6 Experimental Results

For the seller side, we generate product descriptions by several LLMs. For the user side, we simulate consumer actions with our previous Oracle model (Qwen-2 72B-FT). We evaluate the average conversion rate (CVR) for 8 different categories for each model. In e-commerce settings, the higher the CVR, the better the model aligns with consumer preferences and facilitates effective product recommendations. The average e-commerce CVR is between 1% and 4%, but can vary significantly depending on the industry, business model, and other factors.

We follow the best practice of (Ethayarajh et al., 2024) and set $\lambda_D = \lambda_U = 1$ in Equation 3 to train our Qwen-KTSA model. Results show that contents generated by Qwen-KTSA outperforms in 7 of 8 categories, demonstrating its ability to understand latent user needs and make accurate product selections.

7 Conclusion and Limitations

In this work, we propose to apply Kahneman-Tversky Optimization (KTO), a novel training strategy into the LLM-based product recommendation scenario. It departs from traditional Maximum Likelihood Estimation (MLE) objectives commonly used in LLM-based generation. By directly optimizing for user utility rather than surface-level linguistic alignment, our approach offers a new perspective for training language models to better capture users' latent goals and decision-making signals in recommendation contexts. Moreover, we present a self-alignment framework where a content generator interacts with a user simulator, allows the model to iteratively improve its outputs through internal feedback loops.

Nonetheless, this work has several limitations. While simulation provides a scalable way to approximate user preferences, it remains an abstraction that may not fully reflect human complexity. Conducting large-scale human evaluations with fine-grained utility judgments is costly and difficult to generalize. Furthermore, marketing, consumption, and recommendation behaviors are highly contextual and multifaceted. Our current focus on aligning product descriptions with inferred user needs addresses only one limited aspect of a much broader problem space.

References

2024. Amazon kdd cup 2024: Multi-task online shopping challenge for large language models. https://amazon-kddcup24.github.io/. Accessed: 2025-05-06.

Chris Deotte, Ivan Sorokin, Ahmet Erdem, Benedikt Schifferer, Gilberto Titericz Jr, and Simon Jegou. 2024. Winning amazon kdd cup'24.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Wei Jin, Haitao Mao, Zheng Li, Haoming Jiang, Chen Luo, Hongzhi Wen, Haoyu Han, Hanqing Lu, Zhengyang Wang, Ruirui Li, et al. 2024a. Amazon-M2: A multilingual multi-locale shopping session dataset for recommendation and text generation. *Advances in Neural Information Processing Systems*, 36.

Yilun Jin, Zheng Li, Chenwei Zhang, Tianyu Cao, Yifan Gao, Pratik Jayarao, Mao Li, Xin Liu, Ritesh Sarkhel, Xianfeng Tang, et al. 2024b. Shopping mmlu: A massive multi-task online shopping benchmark for large language models. *arXiv preprint arXiv:2410.20745*.

Seungjae Jung, Gunsoo Han, Daniel Wontae Nam, and Kyoung-Woon On. 2024. Binary classifier optimization for large language model alignment. *ArXiv*, abs/2404.04656.

Akira Kasuga and Ryo Yonetani. 2024. Cxsimulator: A user behavior simulation using llm embeddings for web-marketing campaign assessment. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 3817–3821. ACM.

Tongyoung Kim, Jeongeun Lee, Soojin Yoon, Sunghwan Kim, and Dongha Lee. 2025. Towards personalized conversational sales agents : Contextual user profiling for strategic action.

Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. 2023. Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce.

OpenAI. 2023. Gpt-4 technical report. *PREPRINT*.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*.

Bo Peng, Xinyi Ling, Ziru Chen, Huan Sun, and Xia Ning. 2024. ecellm: Generalizing large language models for e-commerce from large-scale, high-quality instruction data. *arXiv preprint arXiv:2402.08831*.

Qwen. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Chandan K Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022. Shopping queries dataset: A large-scale esci benchmark for improving product search. *arXiv preprint arXiv:2206.06588*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Amos Tversky and Daniel Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5:297–323.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.