Combining A/A Data and LLMs for Improved Content Evaluation

Shiv Shankar
University of Massachusetts
USA

Madalina Fiterau
University of Massachusetts
USA

Abstract

A/B testing to evaluate user preferences and engagement is a cornerstone of the modern digital landscape. However, in the current era, the feedback cycle is considerably shortened while the experimentation space expands significantly, necessitating novel and efficient ways to assess user engagement. A/A testing, which compares identical content variants, offers a complementary approach by establishing baselines for engagement metrics and identifying natural variability in user behavior. However, A/A tests inherently lack paired samples, limiting their direct applicability to standard preference alignment methods, which require positive and negative samples for the same context. To address this gap, we propose a novel utility theory framework that enables the integration of unpaired A/A data into content evaluation systems. By translating Large Language Model (LLM) rewards into a utility framework, our approach allows for the incorporation of A/A test results, into predictive models.

CCS Concepts

• **Applied computing** → **Electronic commerce**; **Marketing**; • **Computing methodologies** → *Machine learning algorithms*; • **General and reference** → *Evaluation*; *Experimentation.*

1 Introduction

Widespread adoption of mobile devices and increased internet access has led to a significant increase in digital content consumption. To maximize customer engagement, businesses constantly aim to optimize the content and user experience. For example, news media industries constantly strive to come up with attractive headlines and cover images [8] to drive customer engagement. The standard practice to find attractive headlines is to use A/B testing. However, this is inefficient for applications surrounding social-media, news and related sectors; as news and trends have short lifetimes and might become irrelevant by the time a standard A/B test finishes. Thus, in industries, where newer content constantly comes up, there is a great need for more-efficient engagement evaluation.

One additional source of data for this purpose can come from A/A data. A/A testing involves comparing two identical versions of content (A vs. A) to establish a baseline for variability and noise in engagement metrics. By analyzing A/A data, organizations can better understand the natural fluctuations in user behavior and engagement, which helps in distinguishing true performance differences in A/B tests from random variations. While data from such runs are used primarily to test the system and improve statistical significance, such data can hold additional signal which often gets discarded in favor of paired results from an A/B test. Often, however such signal is present in a complex manner, distributed across different contexts. This is where Large Language Models (LLMs), can provide the necessary representational capacity to utilize such signal. LLMs have demonstrated the ability to mimic human preferences and behavior in a variety of consumer research tasks [6, 18]; and in this work we raise the natural question is "Can A/A data along with LLMs can be improve content rating models?".

Summary. In this paper, we explore fine-tuning LLMs for A/B testing with additional A/A data. Since standard preference learning methods cannot leverage unpaired (A/A) data, we propose a modified approach inspired by utility theory that can utilize A/A data to improve performance. For concreteness, we will consider writing headlines for articles as our running example. As such we will use the terms content/article/prompt and the terms treatment/headline interchangeably. Our experiments suggest that with suitable training LLMs can leverage such data to improve performance, while smaller models like BERT are less effective.

2 Preliminaries and Related Work

2.1 Learning from A/B Tests

Following Kaufmann et al. [14] we treat the problem of A/B testing in a preference learning framework and follow standard notation from literature[29]. The language model is considered as a policy function π that observes a prompt x and produces a textual response a by sampling from a distribution $y \sim \pi(\cdot \mid x)$. We are given a dataset $\mathcal{D}_{\text{pref}} = (x, a^+, a^-)$ consisting of prompts and labeled response pairs. Here, a^+ represents the positive response, and a^- represents the negative response. For example, in A/B testing different summaries or headlines for given content, the preference data is collected by exposing incoming traffic to one of two possible treatments (A or B). The resulting engagement, measured via metrics such as clicks, screen time, or another chosen metric, is monitored. The option with higher engagement is taken as the positive sample a^+, while the other is taken as negative sample a^-.

Offline RLHF. [7, 24, 29] addresses the challenge of aligning a policy network using $\mathcal{D}_{\text{pref}} = \{(x, a^+, a^-)\}$. Given a context or prompt x, a pair of outputs is sampled from $\pi_{\text{ref}}(\cdot \mid x)$ and then ranked based on a preference function, which is often determined by human annotations. RLHF methods [7, 24] aim to learn a policy $\hat{\pi}$ that aligns with this preference data. As in previous work [24], we consider the *Bradley-Terry* model [5] for preferences, though other models could also be applied. Typically, the process begins by estimating a reward function $\hat{r}$ from $\mathcal{D}_{\text{pref}}$ using maximum likelihood estimation (MLE). Once $\hat{r}$ is obtained, reinforcement learning methods like PPO are used to optimize $\hat{r}$, with an additional regularization term

Proceedings of the ACM Conference 2024, April 28-May 2, 2024, Australia ,

$$\hat{\pi} = \underset{\pi \in \Pi}{\operatorname{argmax}} \mathbb{E}_{\pi} \left[\hat{r}(x, y) - \beta \log \frac{\pi(a \mid x)}{\pi_{\text{ref}}(a \mid x)} \right].$$

The optima for the above model is given by the energy model $\hat{\pi} \propto \pi_{\text{ref}} \exp(\hat{r}/\beta)$ [32]. Using this insight the DPO method [27] relies on optimizing π directly via plugging the corresponding implied reward function in MLE objective, leading to :

$$\max_{\pi} \sum_{(x,a^+,a^-) \in \mathcal{D}_{\text{pref}}} \log \sigma \left(\beta \frac{\pi(a^+ \mid x)}{\pi_{\text{ref}}(a^+ \mid x)} - \frac{\pi(a^- \mid x)}{\pi_{\text{ref}}(a^- \mid x)} \right), \quad (1)$$

Other methods like SLIC [31], IPO [1], follow the same insight and use the log-likelihood ratio between the LLM policy and the reference model as the implied reward model for tuning LLMs.

Utility Theory. is a fundamental concept in economics and decision theory, for modeling choices under uncertainty [20]. Given a space $\mathcal{X}$ of outcomes, a utility function $u : \mathcal{X} \to \mathbb{R}$ assigns a real value to a real number to each $x \in \mathcal{X}$. The utility function is meant to capture the decision-maker's preferences, where: $x_1 \geq x_2$ iff $u(x_1) \geq u(x_2)$. Typically u is a positive non-decreasing function when $\mathcal{X} \subseteq \mathbb{R}$.

LLM Alignment is a major research topic with many different methods proposed. Yuan et al. [30] uses ranking loss instead of sigmoid loss. Ethayarajh et al. [12] uses a reference point for computing reward/loss from a sample. Many proposals also use different forms of including using 'chain-of-thought' COT reasoning [25] and using curriculum learning [26]. Our work while learning alignment is more focused on using the LLM as a rater for the treatment arms, instead of specific alignment with human preferences. For example, a model which can rate the treatments correctly may not be good at generating text.

Other Related Work. The connection between stochastic ordering and preference learning has been previously noted [22, 23]. Melnyk et al. [22] propose minimizing the Wasserstein distance between reward functions and link it to stochastic ordering. Other connections between optimal transport-related methods and stochastic ordering have been well studied [10, 15, 16]. Borrowing ideas from prospect theory, the KTO approach of Ethayarajh et al. [12] optimizes the margin between the chosen reward and the average reward of rejected sentences. This is a form of stochastic dominance that focuses on population means.

3 Method

We now present our methodology for incorporating partial results from an experiment, particularly when results from A/A tests are available. An A/A test is a randomized experiment similar to an A/B test. However, unlike an A/B test, an A/A test exposes both the treatment and control groups to the same conditions rather than different ones. The main purpose of an A/A test is typically to validate the experimental setup and estimate variability.

Since in an A/A test we evaluate only one treatment, and an A/A test only gives data in the form (x, a), unlike an A/B test which provides data in the form (x, a^+, a^-). Therefore, in this section, we focus on scenarios where triplets of articles and positive/negative headlines (x, a^+, a^-) are not available. Instead, we assume access to two separate distributions: μ_+, which contains positive pairs (x_+, a^+) representing higher click rate examples that the model should imitate, and μ_-, which contains negative samples (x_-, a^-) associated with lower click rates. These distributions can be derived from an A/A test, though other sources of such data are also possible. For example, marketers often create multiple alternatives before narrowing down to the most promising ones. In such cases, the unselected examples can serve as a potential source of such data.

Consider the problem of comparing two distributions, μ_+ and μ_-. We have access to datasets $\mathcal{D}_+ = (x_+, a^+)$ and $\mathcal{D}_- = (x_-, a^-)$, which are sampled from μ_+ and μ_-, respectively. Here, x_+ and x_- represent features or variables, while a^+ and a^- represent associated outcomes or actions. Ideally, we aim to learn a model or decision rule that favors the positive example distribution μ_+ over the negative example distribution μ_-. However, since the x-values between the two datasets are not shared, it is not possible to compare them at the instance level, as is done in standard preference learning methods like RLHF [7], DPO [27] and IPO [2].

While individual tuples cannot be directly compared, we can approach this problem from a decision-theoretic perspective. In decision theory, preferences are typically defined in terms of the decision maker's utility. Following this idea, we compare the aggregate utilities derived from each dataset. Let $u(r(x, a))$ represent the decision maker's utility function, which assigns a real number to each pair (x, a). Since we want a model that prefers $\mathcal{D}_+/\mu_+$ over $\mathcal{D}_-/\mu_-$; we need to ensure that the expected utility from the dataset μ_+ is greater than the expected utility from μ_-, meaning:

$$\mathbb{E}_{(x,a)\sim\mu_+}[u(x, a)] > \mathbb{E}_{(x,a)\sim\mu_-}[u(x, a)], \quad (2)$$

where the expectation is taken over the distributions μ_+ and μ_-. This comparison would indicate that, on average, the dataset μ_+ yields higher utility than μ_-. However, the true utility function u for the decision-maker is unknown, which making this constraint non-trivial to enforce.

To resolve this, we instead consider the same constraint over all reasonable utility functions u. We restrict our attention to utility functions that are positive, increasing, and concave (which is a reasonable assumption in many economic and decision-making contexts). When u satisfies these conditions, eq. (2) becomes related to the idea of **Stochastic Dominance**, which is a way to compare probability distributions in terms of risk preferences. A distribution μ_+ is said to second-order stochastically dominates μ_- (written as $\mu_+ \underset{\text{SD}}{\geq} \mu_-$), if for all increasing concave utility functions u, the following condition holds:

$$\mu_+ \underset{\text{SD}}{\geq} \mu_- \text{ if } \mathbb{E}_{z\sim\mu_+}[u(z)] \geq \mathbb{E}_{z\sim\mu_-}[u(z)]. \quad (3)$$

Equation 3 simply asserts that the expected utility from μ_+ is higher than μ_-. Intuitively, if one distribution stochastically dominates another, it is preferred by all decision-makers who are risk-averse (i.e., who have concave utility functions).

To apply this to the LLM setting we note that the implicit r used by the trained final LLM is related to the final model's likelihood as in Rafailov et al. [27], Ziebart et al. [32]. Specifically, as the implicit reward given an RLHF model π_θ, is given by $\log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$, we can define distributional preference for π_θ as follows:

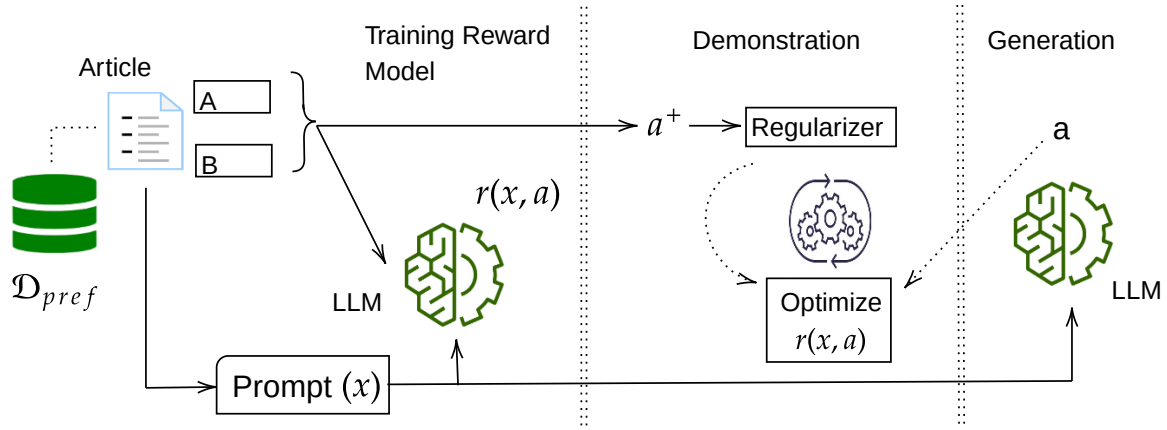

Figure 1: Overview of the proposed generative approach. The reward model r is obtained by tuning an LLM on the preference data $\mathcal{D}_{\text{pref}}$, which consist of tuples of contexts/articles along with two treatments arms (a^+, a^-). Given a prompt x (which includes the context/article along with instructions) the generator LLM produces an output a. The pair x, a^+ is considered as a demonstration for the generator to match and improve using the reward model r.

Definition 1 (Distributional Preference [22]). A policy π_θ prefers distributionally μ_+ on μ_- with respect to a reference policy π_{ref} if:

$$\log \frac{\pi_\theta(a^+|x_+)}{\pi_{\text{ref}}(a^+|x_+)} \underset{\text{SD}}{\geq} \log \frac{\pi_\theta(a^-|x_-)}{\pi_{\text{ref}}(a^-|x_-)}.$$

Remark 1. *Note that in DPO[27] the dataset is paired, and hence we have a^+, a^- for the same x. In the above constraint replaces x^+, x^- are replaced by x in that setting. Thus we get an alternative interpretation of DPO and related methods [2, 31] as imposing the above constraint in the form of imposing a pointwise preference[1] order for a^+ over a^-.*

The expected utility version while theoretically sound is not amenable for direct optimization. As the reward model changes, we would need to learn a new utility function, leading to an minimax optimization problem. Instead we can use the following classic theorem from decision theory connects the utility functions to the problem related to quantiles of the distribution [20]

Theorem 3.1. *(Proposition 6.D.2 in Mas-Colell et al. [20]) Let $F_+(z)$ and $F_-(z)$ represent the cumulative distribution functions (CDFs) of the datasets μ_+ and μ_-, respectively. The condition for μ_+ to stochastically dominate μ_- as in Equation* (3) *is equivalent to:*

$$\int_{-\infty}^{s} F_+(t)\,dt \leq \int_{-\infty}^{s} F_-(t)\,dt \quad \text{for all } s \in \mathbb{R}.$$

This integral condition ensures that, under any concave utility function, the expected utility of μ_+ is greater than or equal to that of μ_-. The integrated quantiles provide a more intuitive way of comparing the distributions, capturing the idea that the "cumulative utility" from μ_+ is at least as large as from μ_- at all points along the distribution. This definition also provides a direct way to both measure and optimize violations via:

$$\mathcal{L}(\mu_+, \mu_-) = \int_{-\infty}^{\infty} h(F_+^{(2)}(t) - F_-^{(2)}(t))\,dt, \tag{4}$$

where $h(.)$ is the 0/1 loss $\mathbb{I}(x > 0)$ and $F^{(2)}$ is the integrated CDF i.e. $F^{(2)}(t) = \int_{-\infty}^{t} F(t)dt$. It is easy to see that L is 0 iff Theorem 3.1 holds, which by Theorem 1 means μ_+ stochastically dominates μ_-.

Since the 0/1 objective has a gradient of 0 almost everywhere, it is not suitable for optimization. Instead, we can replace it with a smoother convex approximation, such as the hinge loss $h(x) = (1-x)_+$ or the logistic loss $h(x) = \log(1+\exp(-x))$. Note that if we scale the input to the logistic loss by β, the expression becomes very similar to the DPO loss objective (Equation (1))[2]. Other objectives like the squared hinge loss has also been suggested [1] in this context for learning preferences.

We propose adding the empirical version of Equation (4) as a minimization objective when tuning the model with additional unpaired data. Let $\psi(a, x) = \frac{\pi(a|x)}{\pi_{\text{ref}}(a|x)}$. The overall objective becomes:

$$\hat{\pi} = \underset{\pi \in \Pi}{\text{argmax}}\, \mathbb{E}_\pi \Big[r_m(x, a) - \beta \log \psi(a, x) + \lambda \mathcal{L}(\mathcal{D}_+, \mathcal{D}_-) \Big], \tag{5}$$

where $\mathcal{L}(\mathcal{D}_+, \mathcal{D}_-)$ is the empirical version of Equation (4). Since this loss involves empirical CDFs, which are weighted sums of step functions, back-propagation through it is challenging. Hence, we rely on ideas from optimal transport [4, 9]. Specifically, Blondel et al. [4] provided a differentiable version of sorting using entropic optimal transport. This can be used to differentiably 'count' the number of values less than [3] t to get a differentiable empirical CDF. From the empirical CDF we can compute the integrated CDF by:

[1] We are not the first to notice this, and Melnyk et al. [22] have suggested such an interpretation as well

[2] We used the scaled sigmoid in our experiments

[3] SoftRank is available in the library torchsort

$$F^{(2)}(t) = \int F(\psi)d\psi \approx \sum_{\psi_{i+1}<t} \frac{F(\psi_{i+1}) + F(\psi_i)}{2}(\psi_{i+1} - \psi_i)$$
$$\approx \sum_{\psi_{i+1}<t} \frac{SoftRank(\psi_{i+1}) + SoftRank(\psi_i)}{2N}(\psi_{i+1} - \psi_i)$$

The values of t for evaluating the integral are chosen based on quantiles of all ψ values. [4].

Regularizing Objectives. Compared to the standard RLHF framework of [24], working with unpaired preferences is more subtle. RLHF is also known to overfit the reward model, but without paired data, the reward model is even more unconstrained and can lead to massive overfitting. To prevent this we constrain the model, to say even closer to the base SFT model than allowed under DPO/RLHF. To accomplish this we include an additional term of $\frac{\pi(a|x)}{\pi_{\text{ref}}(a|x)}$ as a regularizer in the objective. This terms more strongly penalizes deviations of π from π_{ref} than just the KL divergence. An astute reader might also note that this term is equivalent to regularizing with the order-2 Tsallis divergence. Thus we get the following maximization objective:

$$\hat{\pi} = \underset{\pi\in\Pi}{\text{argmax}}\, \mathbb{E}_\pi \Big[r_m(x,a) - \beta \log \psi(a,x) - \beta\psi(a,x) + \lambda \mathcal{L}(\mathcal{D}_+, \mathcal{D}_-) \Big], \quad (6)$$

where β, λ are hyperparameters. Since the data source is limited and training a full LLM requires significant compute resources, we use LORA [13] based training for the above objective using the PEFT [19] library.

4 Experiments

Dataset. We experiment with a public datases obtained from real-life A/B testing scenarios [21]. The data consists of several versions of headlines created by an editorial teams for various articles. Each user was exposed to only one of these headlines article pair, and the clicks were recorded for each pair. We remove any duplicates or image linked headlines (so we focus only on textual content). We only considered text only content and restricted to treatments that have statistically different CTRs (at p=0.10).

Baselines and Methodology. Recent research [3, 17] suggests that using an LLM based embedding model is a strong performer, especially on treatment rating [28]. Thus, we include GPT3 embedding as a baseline. Since this is a classifier, it can only use A/B data. In addition to Llama we also train our proposed approach of using AA data with a smaller LMs, specifically BERT [11]. Finally, to assess the efficacy of having the additional A/A data, we train the model with only A/B data, and labeled GenLLMAB. We also train BERT with the distributional loss (called BERT+AA). Since our dataset does not inherently include A/A data for evaluation, we simulate such a setting by providing only a fraction of the full training data as A/B data. For the remaining portion of the training data, we randomly select one treatment as data from an A/A test, ignoring all other treatments. This process naturally creates dataset splits with p% A/B data and $(100-p)$% A/A data.

[4]In practice we also regularized with $|\mathbb{E}_{\mathcal{D}_+}[\psi] - \mathbb{E}_{\mathcal{D}_-}[\psi]|$

Results. In Figure 2, we plot the accuracy of the different models across varying availability of unpaired data. We can see that when we have only a little A/B data, the generative approach can overfit, and using embeddings is a better approach. However when augmented with the additional data, the model can do better by almost 5% accuracy points in the low data regime. From the figure we can see also that BERT based model does not improve much with our proposed training . This is because to generalize across different context using the distributional loss needs a stronger model.

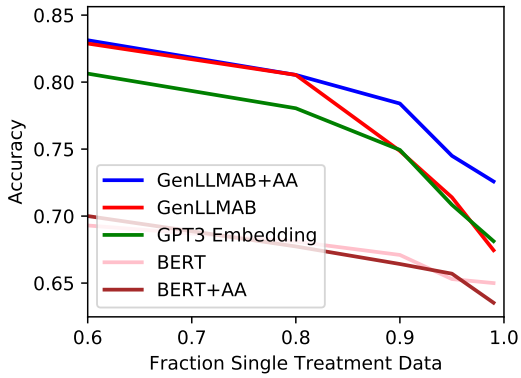

Figure 2: Plot of accuracy of different models as the amount of A/B data decreases. The x-axis is the fraction of A/A (or unpaired) data.

5 Conclusion

In this work, we address the limitations of traditional A/B testing in fast-paced industries by proposing a novel framework to incorporate unpaired A/A data into content evaluation systems. Our experiments demonstrate that Large Language Models (LLMs), when suitably trained, can effectively leverage A/A data to improve predictive performance, while smaller models like BERT struggle to achieve similar gains. This highlights the importance of model scale and capability in handling unpaired datasets. Furthermore, in low-data regimes, our approach shows significant promise, with up to a 5% increase in predictive power when A/A data is utilized. By introducing a utility-theoretic framework, we provide a principled method for translating LLM rewards into distributional utilities, enabling the integration of unpaired data into preference alignment and content optimization tasks. Future work could explore the application of this framework to broader contexts, such as personalized recommendations and adaptive user interfaces.

References

[1] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4447–4455.

[2] Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A General Theoretical Paradigm to Understand Learning from Human Preferences. arXiv:2310.12036 [cs.AI] https://arxiv.org/abs/2310.12036

[3] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961* (2024).

[4] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. 2020. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*. PMLR, 950–959.

[5] Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39 (1952).

[6] James Brand, Ayelet Israeli, and Donald Ngwe. 2023. Using gpt for market research. *Available at SSRN 4395751* (2023).

[7] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences.

[8] Anna Coenen. 2019. How The New York Times is Experimenting with Recommendation Algorithms. https://nyti.ms/3No7dYS

[9] Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. 2019. Differentiable ranking and sorting using optimal transport. *Advances in neural information processing systems* 32 (2019).

[10] Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. 2018. An optimal transportation approach for assessing almost stochastic order. *The Mathematics of the Uncertain: A Tribute to Pedro Gil* (2018), 33–44.

[11] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[12] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306* (2024).

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

[14] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. 2014. On the complexity of A/B testing. In *Conference on Learning Theory*. PMLR, 461–481.

[15] Jakwang Kim, Young-Heon Kim, Yuanlong Ruan, and Andrew Warren. 2024. Statistical inference of convex order by Wasserstein projection. *arXiv preprint arXiv:2406.02840* (2024).

[16] Lasse Leskelä and Matti Vihola. 2017. Conditional convex orders and measurable martingale couplings. (2017).

[17] Chaofan Li, MingHao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. 2024. Making text embedders few-shot learners. *arXiv preprint arXiv:2409.15700* (2024).

[18] Peiyao Li, Noah Castelo, Zsolt Katona, and Miklos Sarvary. 2024. Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis. *Marketing Science* (2024).

[19] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. https://github.com/huggingface/peft.

[20] Andreu Mas-Colell, Michael Whinston, and Jerry Green. 1995. *Microeconomic theory*. Oxfors University Press.

[21] J Nathan Matias, Kevin Munger, Marianne Aubin Le Quere, and Charles Ebersole. 2021. The Upworthy Research Archive, a time series of 32,487 experiments in US media. *Scientific Data* 8, 1 (2021), 195.

[22] Igor Melnyk, Youssef Mroueh, Brian Belgodere, Mattia Rigotti, Apoorva Nitsure, Mikhail Yurochkin, Kristjan Greenewald, Jiri Navratil, and Jerret Ross. 2024. Distributional Preference Alignment of LLMs via Optimal Transport. *arXiv preprint arXiv:2406.05882* (2024).

[23] Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. 2023. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886* (2023).

[24] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155

[25] Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733* (2024).

[26] Pulkit Pattnaik, Rishabh Maheshwary, Kelechi Ogueji, Vikas Yadav, and Sathwik Tejaswi Madhusudhan. 2024. Enhancing Alignment using Curriculum Learning & Ranked Preferences. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 12891–12907.

[27] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: your language model is secretly a reward model *(NIPS '23)*.

[28] Shiv Shankar, Ritwik Sinha, and Madalina Fiterau. 2024. On LLM Augmented AB Experimentation. In *Causality and Large Models @NeurIPS 2024*. https://openreview.net/forum?id=dgeWznoY8h

[29] Chenlu Ye, Wei Xiong, Yuheng Zhang, and Tong Zhang. 2024. Online Iterative RL from Human Feedback with General Preference Model. arXiv:2402.07314

[30] Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2024. RRHF: Rank responses to align language models with human feedback. *Advances in Neural Information Processing Systems* 36 (2024).

[31] Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. 2023. Calibrating Sequence likelihood Improves Conditional Language Generation. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=0qSOodKmJaN

[32] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. 2008. Maximum entropy inverse reinforcement learning.. In *Aaai*, Vol. 8. Chicago, IL, USA, 1433–1438.