Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Sister Conferences Best Papers Track

Mechanism Design for Large Language Models (Extended Abstract)*

Paul Dütting[1] , **Vahab Mirrokni**[1] , **Renato Paes Leme**[1] , **Haifeng Xu**[1,2] , **Song Zuo**[1]
[1]Google Research
[2]University of Chicago
{duetting,mirrokni,renatoppl,szuo}@google.com, haifengxu@uchicago.edu

Abstract

We investigate auction mechanisms for AI-generated content, focusing on applications like ad creative generation. In our model, agents' preferences over stochastically generated content are encoded as large language models (LLMs). We propose an auction format that operates on a token-by-token basis, and allows LLM agents to influence content creation through single dimensional bids. We formulate two desirable incentive properties and prove their equivalence to a monotonicity condition on output aggregation. This equivalence enables a second-price rule design, even absent explicit agent valuation functions. Our design is supported by demonstrations on a publicly available LLM.

1 Introduction

In the current web ecosystem, auctions are the primary mechanism used to decide which ads (and commercial content more broadly) are displayed to users [Edelman *et al.*, 2007; Varian, 2007]. In these auctions, advertisers bid for the opportunity to display their ad creatives alongside organic content. Many of the web formats such as text, banners, video, apps, ... have their own subtleties which led to the development of new auction tools to handle them. Our goal in this paper is to investigate auction mechanisms to support the emerging format of AI-generated content. More specifically, we explore the use of auctions as a tool for influencing the output of large language models (LLMs) (e.g., [Brown *et al.*, 2020]).

We consider a situation where a certain space in the web (which could be a part of a webpage, an UI element of an AI-chatbot, the dialog of a certain character in a video or a game, etc.) is designated for commercial content and different advertisers can bid to influence the content in that space. Each advertiser has an LLM that can generate content for that space, and is willing to pay a certain amount of money for the right to have their content displayed. A simple design is to collect bids from advertisers and let the highest bidder choose whatever content they wish to publish in that space. While simple, this design does not exploit the flexibility of LLMs which is to combine different concepts in a creative way.

Consider this example. First, we ask an LLM to produce different ads for the fictitious Stingray Resort and the equally fictitious Maui Airlines:

- "*Experience the magic of Hawaii at Stingray Resort, where stunning views, luxurious accommodations, and endless activities await. Book your stay today and create unforgettable memories in the heart of paradise*."
- "*Fly to Hawaii with Maui Airlines and experience the beauty of the Aloha State. We offer affordable flights to all the major islands, so you can start your Hawaiian vacation sooner. Book your flight today and let the island spirit take over!*"

For that use case, however, the LLM is flexible enough to produce a joint ad for both:

- *"Fly to paradise with Maui Airlines and experience the magic of Hawaii at Stingray Resort. Stunning views, luxurious accommodations, and endless activities await. Book your dream vacation today and create unforgettable memories."*

One can envision an auction mechanism that allows both Stingray Resort and Maui Airlines to submit their LLMs and bids, with these inputs determining their prominence in the final outcome.[1]

1.1 Unique Challenges

LLMs [Brown *et al.*, 2020; Thoppilan *et al.*, 2022; Google *et al.*, 2023] are an emerging technology with new and unconventional aspects, many of which have direct implications to auction design (e.g., how preferences are represented/expressed). Our goal is to identify some of the key challenges and take a first step in designing mechanisms to address them:

*Original paper: Paul Dütting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. In *Proceedings of the ACM Web Conference 2024*, pages 144–155, 2024. The work of Haifeng Xu was done as Visiting Faculty at Google Research.

[1]While this work's main focus is to create ad creatives that merge content from different advertisers, our designed auction mechanism for merging LLM outputs could also be used in other contexts.

- **Modelling and Expressing Preferences.** Auction theory typically models preferences via value functions that assign a value to each outcome. LLMs, however, as *generative models*, do not directly assign values. Instead, they succinctly encode preferences over outcomes within a stateless neural network model that predicts continuation probabilities.
- **Necessity of Randomization.** LLMs crucially rely on randomization. When forced to output tokens deterministically, LLMs often have a worse performance compared to situations that sample from a distribution (see, e.g., [Holtzman *et al.*, 2019], for a performance comparison of different decoding strategies). Therefore, an auction that aggregates LLM outputs should preferably also output distributions.
- **Technical Compatibility.** Auction solutions should be compatible with current LLM technology, utilizing readily available information and integrating seamlessly. Ideally, the allocation and payments should be obtained from simple manipulations of the LLM outputs.
- **Computational Efficiency.** LLM models are expensive to query, so the auction computation should not add too much overhead. In particular, auctions should not increase the number of calls to inference the models beyond the minimum necessary.

1.2 Our Contributions

The Token Auction Model. Our first contribution is a formalism ("The Token Auction Model") for studying this problem. *Tokens* are the units making up sentences and paragraphs.[2] Examples of tokens include (sub-)words, symbols, numbers, and special tokens indicating the beginning and ending of the text. In particular, any piece of text (potentially incomplete) can be represented as an array of tokens, and any array of tokens also encodes a piece of text.

One salient feature of the state-of-the-art LLMs is that they are stateless, i.e., they maintain no internal memory or state. Instead, they simply map a prefix string to a distribution over the next token. The output is then created in an autoregressive manner. Given an input prompt, the output is generated by repeatedly feeding the current sequence of tokens into the LLM, sampling a continuation token, and appending it to the sequence of tokens.

The proposed *token auction* operates on a token-by-token basis, and serves to aggregate several LLMs to generate a joint output. We assume the designer has access to algorithmic LLM agents represented by their respective text generation functions (the functions that map a sequence of tokens to a distribution over the next token). In addition, we allow each LLM agent to submit a single dimensional bid. The auction output will be an aggregated distribution together with a payment rule that defines payments for each agent.[3]

This approach may seem counterintuitive initially, as advertisers typically focus on the final generated text rather than individual word choices. This seems to suggest a dynamic planning of the generated token sequence. However, existing LLMs do not reason about full pieces of text, nor do they plan ahead; instead, their preferences are expressed as desired distributions over merely the next token. In other terms, we can think of an LLM as a succinct distillation of an agent's complex combinatorial preferences over sequences of tokens into a generative token-by-token model.[4]

The problem of aggregating LLMs forces the designer to understand the preferences of the agents away from the distilled LLM. This appears to be a very difficult problem. Specifically, we believe it is implausible or at least impractical to assume an individual agent can meaningfully manipulate the distribution over tokens at any given stage, to direct the produced text to a more preferred one. Our auction formulation seeks to strike a balance: By truthfully revealing the LLM to the designer, the agent gives the auction mechanism a hint as to what their preferred distribution is. The bids, in turn, can be used to tradeoff between agents, and in particular help the designer determine their relative weights.

Simple and Robust Token Auctions. Motivated by the challenges in modeling agents' preferences over generated distributions, we take a robust design approach aiming for token auctions that provide desirable incentive properties, while imposing minimal assumptions on the agents' preferences over distributions.

Specifically, we model agents' preferences as entailing partial orders over distributions. Based on this partial preference order[5], we formulate two desirable incentive properties, which we consider minimal requirements:

- *Payment monotonicity*: Given two different bids by the same agent, a final distribution is closer to the desired distribution if and only if the payment is higher.
- *Consistent aggregation:* If for two different bids of the same agent, the final distribution is closer to the preferred distribution for some bids of the other agents, then it should be so for all bids of the other agents.

We show that any mechanism with these two properties is *strategically equivalent* to a mechanism that satisfies a monotonicity requirement on the distribution aggregation function.

We then investigate whether it is possible to equip such distribution aggregation functions with payment rules that satisfy additional incentive properties. Specifically, we investigate whether such aggregation rules admit an analogue of the *second-price payment rule*. In the single-item second-price (or Vickrey) auction [Vickrey, 1961], the payment corresponds to the critical bid where an agent transitions from losing to winning. To port this notion to our setting, we show that under robust preferences, any monotone aggregation rule can be written as a distribution over deterministic allocations from bids to tokens such that there is a critical bid where the

[2] More generally, one can consider tokens forming parts of images [Ramesh *et al.*, 2021; Yu *et al.*, 2022] and videos [Sun *et al.*, 2019]. For the purpose of this paper, we stick with text generation.

[3] See our discussion later this section on the rationale of the indirect mechanism formulation.

[4] See our discussion in the original paper for additional support for the stateless approach.

[5] Partial orders are more general than total orders, and hence our key results apply to any complete preference order model.

allocation transitions from a less preferred to a more preferred token. Such a critical bid then serves as a natural candidate for a payment rule. This hence leads to an analogue of the second-price auction for our token auction model that only requires ordinal preferences. The resulting class of auctions is applicable whenever the agent valuations are compatible with the partial order, yielding robust incentives for all of these.

Designing Aggregation Functions. We then move to designing concrete aggregation functions. Our approach considers aggregated loss functions inspired by state-of-the-art LLM training, and derives optimal distribution aggregation functions that minimizes such aggregated loss functions.

We focus on specific forms of aggregated loss functions based on KL-divergence, a commonly used loss function in LLMs. We consider two natural formulations inspired by current LLM training, and show that the corresponding optimal aggregation rules are the weighted (log-space) convex combination of the target distributions from all participants.

The linear and log-linear aggregation rules we identify have different pros and cons. Both share the advantage that they are optimal for the respective aggregated loss functions. The linear rule turns out to be monotone with respect to robust preferences, and is therefore compatible with the robust incentives approach. However, the log-linear rule is not.

Demonstration. We conclude with demonstrations to support our token auction formulation, obtained by prompt-tuning of a publicly available LLM. A two-advertiser demonstrative example is considered, under both the linear and log-linear aggregation rules. We show how the combined output varies as a function of $\lambda = {}^{b_1}/_{(b_1+b_2)}$, where b_1 and b_2 are the advertisers' bids. Both approaches lead to meaningful and interpretable texts that smoothly transition from favoring one to favoring another advertiser, with a joint ad produced for intermediate values of λ.

Discussion/Design Choices. An alternative to our approach of designing an *indirect mechanism* would be to aim for a *direct mechanism*. Such a mechanism, instead of asking agents for a scalar bid along with query access to the agents' LLMs, would elicit the agents' full preferences directly. However, this appears unrealistic in our new domain due to multiple reasons: (1) Allocation outcomes in our setting are a high-dimensional distribution, whereas a classic mechanism's allocation is typically a subset of items, and often a single item in tractable setups. (2) While it is reasonable in the classic setup to elicit a valuation for an item or a subset of items, it does not appear realistic to elicit a high-dimensional utility function over all possible token distributions. (3) Eliciting full preferences over any token distribution would require solving a problem that is strictly harder than what current LLMs are trained to do (namely, merely output the most preferred distribution). This level of complexity might go beyond current technological capabilities and would likely be computationally inefficient.

1.3 Additional Related Work

To the best of our knowledge, the exact research question and our approaches here have not been previously studied. However, our work is indeed connected to a few lines of research.

Related LLM Research. Our work shares some similarities with the literature on fine-tuning LLMs, with reinforcement learning from human feedback (RLHF) as a representative approach [Wei *et al.*, 2021; Bakker *et al.*, 2022; Ouyang *et al.*, 2022; Bai *et al.*, 2022]. At a high level, fine-tuning and RLHF seek to align a generally pre-trained LLM with certain desirable behaviors. This is in spirit analogous to our goal of designing LLMs to better align with a group of agents' overall preferences. However, our research challenges and methods are both different from those in the fine-tuning literature. Specifically, fine-tuning refines the underlying model's parameters whereas our approach is one-layer up and directly aggregates the token distributions from multiple models. The main challenge we address is the potential incentive misalignment while eliciting LLM agents' preferences, whereas human labelers or other models that generate reward feedback for RLHF are assumed to be genuine and do not misrepresent their own preferences.

The literature on in-context learning [Brown *et al.*, 2020; Wei *et al.*, 2022; Wei *et al.*, 2023] is similar to us in the sense that this approach also does not change the model parameters. A main difference to our work is that this literature seeks to influence token distributions by conditioning on better-generated prefix contexts, whereas we directly aggregate distributions from multiple LLM agents.

Connections in Mechanism Design. Our work is related to the literature on (combinatorial) public projects [Papadimitriou *et al.*, 2008; Dughmi, 2011]. The connection is that one can view the output of the aggregated LLM in our situation as a public project that benefits the agents to different degrees. Similar to these earlier studies, a core challenge in our problem is to elicit preferences about the public project from unknown agents. However, the design problem in our case is fundamentally different — we choose a high-dimensional distribution from an $\mathbb{R}^T$ space with only partial knowledge about agents' preferences, whereas previous work has focused on the problem of choosing from a discrete (often exponentially large) set with clear agent valuation functions [Papadimitriou *et al.*, 2008; Dughmi, 2011].

Another related stream of work includes [Freeman *et al.*, 2019; Goel *et al.*, 2019], which studies the problem of truthfully aggregating budget proposals. Their mechanisms output a distribution over budgets that best serves the population, just like our mechanisms output distributions over tokens. However, the objectives and techniques between our work and theirs are both different. First, their problem is mechanism design without money, whereas our problem has monetary transfers involved. A direct consequence of this first difference is that their mechanisms will treat every participant with equal weight, whereas the weights of our participants are determined by their bids. Second, the research on truthful budget proposal aggregations typically assumes explicit valuation functions (e.g., l_1 distance between preferred and output distributions), under which the VCG mechanism is truthful. Their main research question hence is to study additional properties of the mechanisms such as Pareto-efficiency and certain fairness properties [Freeman *et al.*, 2019]. Assuming such an explicit valuation function does not appear

λ	Linear aggregation function q_{KL}	Log-linear aggregation function $\bar{q}_{\mathsf{KL}}$
1	**Alpha Airlines**: Your ticket to paradise.	
0.75	**Alpha Airlines**: Fly to Hawaii and experience the beauty of the islands with aloha. **Beta Resorts**: Stay at our resorts and enjoy the best of Hawaii.	Feel the magic of Hawaii with a flight on **Alpha Airlines**, now offering 20% off all flights when you book with us today!
0.6	**Alpha Airlines** flies you to Hawaii, where you can enjoy a week-long stay at the **Beta resort** for just $1000.	Experience the magic of Hawaii with a flight on **Alpha Airlines**, now offering 20% off all flights when you book with us today!
0.55	**Alpha Airlines** flies you to Hawaii, where you can enjoy a week-long stay at the **Beta resort**.	Escape to the tropical paradise of Hawaii with [**Alpha Airlines**]!
0.5	**Alpha Airlines** flies you to Hawaii, where you can enjoy a beautiful sunset on the beach. Stay 3 nights and get the 4th free at the **Beta Resort**.	Experience the magic of Hawaii with a stay at the luxurious [**Beta Resort**] and a refreshing flight on [**Alpha Airlines**].
0.45	Fly **Alpha Airlines** to sunny Hawaii and enjoy the secluded beaches and private lagoons of the Royal Hawaiian **Beta Resort**.	Experience the magic of Hawaii with a stay at the luxurious [**Beta Resort**] and a special flight offer from [**Alpha Airlines**].
0.4	Fly **Alpha Airlines** to sunny Hawaii and enjoy the first-class treatment that awaits you at **Beta Resort**, all for one low price.	Experience the magic of Hawaii at the [**Beta Resort**], where you'll feel like you're in a tropical paradise.
0.25	Experience the magic of Hawaii at the **Beta Resort**, where the sun shines brighter and the waves crash louder — book your stay today with our exclusive 20% off discount!	Experience the magic of Hawaii at the **Beta Resort**, where you'll be pampered like royalty and surrounded by breathtaking beauty.
0	Hawaii's **Beta Resort**: a paradise where the sun shines brighter, the waves sing sweeter, and the sand feels softer.	

Table 1: Text generation from two aggregation functions with different $\lambda = b_1/(b_1 + b_2)$.

realistic in our problem, so our core research question is to design robust mechanisms that enjoy good incentive properties simultaneously for a broad range of valuation functions.

From this perspective, our work also bears some similarity to the rich literature on robust mechanism design. Most of this literature still assume existence of value functions with uncertainty modeled by Bayesian beliefs or in a max-min sense [Bergemann and Morris, 2005; Bergemann and Morris, 2012; Roughgarden and Talgam-Cohen, 2016; Carroll, 2015; Dütting *et al.*, 2019]. However, assuming such a valuation function over tokens or their distributions does not appear realistic in creatives generation, thus our model is more similar to a worst-case style consideration during which we only assume partial ("obvious") preferences.

Follow-Up Work. Several papers follow-up on our work, by studying mechanism design problems for LLMs. [Dubey *et al.*, 2024] consider bidders that bid for placement of their content within a summary generated by an LLM. [Soumalias *et al.*,] design a truthful mechanism that generates several samples from a reference LLM, and incentivizes bidders to truthfully reveal their preferences. [Mordo *et al.*, 2024] consider sponsored question answering, in which an organic answer to a search query is fused with an ad to create a sponsored answer, and advertisers bid on the sponsored answers.

2 Demonstration

We implement the aggregation functions we proposed and discuss the examples they produce. Off-the-shelf LLMs generate full text passages. In our case, we need to peek at the internal states of LLMs (the probability distributions over tokens) at each token generation stage. Therefore, we use a custom version of the Google Bard model with a modified inference method that allows access to the token distributions.

Starting from the same base model, we simulate customized LLMs for different agents by agent specific prompt-tuning. A key advantage of simulating LLM agents with different prompts is the ability to use a single LLM, making multiple queries with different prompts instead of serving multiple LLMs concurrently.

2.1 Setups

We illustrate our method with a co-marketing example here (see original paper for a competing brands example), where two agents would like to advertise for their brands, "Alpha Airlines" and "Beta Resort" respectively, regarding a shared topic "Hawaii." We intentionally choose fictitious brands in order to avoid the model directly retrieving any existing ads. We use the brand names "Alpha" and "Beta" that do not have strong meanings to minimize any potential hallucination, as we are using a common purposed LLM that is not optimized for our task. Each agent is given the following prompt:

> "You are an expert of writing texts that naturally combines two ads together. Your choice of words and sentences is full of artistic flair.
> Write a one-sentence ad for __________."

Agent A uses *"a flight to Hawaii using [**Alpha Airlines**]"* to fill the blank, while agent B uses *"a vacation in Hawaii at the [**Beta Resort**]"*. The first two sentences in the prompt aim to improve the quality of the ad generation through *assigning roles* (see, for example, [Wu *et al.*, 2023]). A natural question is whether the proposed method can adjust the combining strategy according to the context. Since in both the linear aggregation rule q_{KL} and the log-linear aggregation rule $\bar{q}_{\mathsf{KL}}$, there is only one degree of freedom, we parameterize the response by $\lambda = b_1/(b_1+b_2)$.

2.2 Results

The results for the co-marketing example are listed in Table 1, where from top to bottom, the value of λ decreases from 1 to 0. As we can see for both aggregation functions, the generated texts roughly follow the pattern of "only Alpha Airlines" $\rightarrow$ "both Alpha Airlines and Beta Resort" $\rightarrow$ "only Beta Resort" when λ decreases. This is expected, as λ going from 1 to 0 corresponds to b_2 increasing from 0 to ∞ with b_1 fixed. The thresholds of pattern changes are 0.75 and 0.4 for the linear aggregation, and 0.5 and 0.45 for the log-linear aggregation. We emphasize that the example is generated with a general purposed LLM, and it is reasonable to believe that the performance can be improved with proper fine-tuning for the specific task at hand.

Acknowledgments

We thank Dirk Bergemann, Marina Halac, Philipp Strack, Elliot Lipnowski, Yang Cai, Vasilis Syrgkanis, Negin Gorezaei, Ido Cohen, Yoav Nagel, Yael Shemesh as well as the participants of the Yale Economics Seminar, the Stanford MS&E Seminar, and the WWW 2024 conference for invaluable comments and suggestions.

We are especially grateful to Yong Cheng from Google DeepMind for his expert guidance on the LLM-related details and literature.

References

[Bai *et al.*, 2022] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862, 2022.

[Bakker *et al.*, 2022] Michiel A. Bakker, Martin J. Chadwick, Hannah Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt M. Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. In *NeurIPS 2022*, pages 38176–38189, 2022.

[Bergemann and Morris, 2005] Dirk Bergemann and Stephen Morris. Robust mechanism design. *Econometrica*, pages 1771–1813, 2005.

[Bergemann and Morris, 2012] Dirk Bergemann and Stephen Morris. *Robust mechanism design: The role of private information and higher order beliefs*, volume 2. World Scientific, 2012.

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. pages 1877–1901, 2020.

[Carroll, 2015] Gabriel Carroll. Robustness and linear contracts. *American Economic Review*, 105(2):536–563, 2015.

[Dubey *et al.*, 2024] Avinava Dubey, Zhe Feng, Rahul Kidambi, Aranyak Mehta, and Di Wang. Auctions with LLM summaries. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 713–722, 2024.

[Dughmi, 2011] Shaddin Dughmi. A truthful randomized mechanism for combinatorial public projects via convex optimization. In *EC 2011*, pages 263–272, 2011.

[Dütting *et al.*, 2019] Paul Dütting, Tim Roughgarden, and Inbal Talgam-Cohen. Simple versus optimal contracts. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 369–387, 2019.

[Edelman *et al.*, 2007] Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1):242–259, 2007.

[Freeman *et al.*, 2019] Rupert Freeman, David M Pennock, Dominik Peters, and Jennifer Wortman Vaughan. Truthful aggregation of budget proposals. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 751–752, 2019.

[Goel *et al.*, 2019] Ashish Goel, Anilesh K Krishnaswamy, Sukolsak Sakshuwong, and Tanja Aitamurto. Knapsack voting for participatory budgeting. *ACM Transactions on Economics and Computation*, 7(2), 2019.

[Google *et al.*, 2023] Google, Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report, 2023.

[Holtzman *et al.*, 2019] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2019.

[Mordo *et al.*, 2024] Tommy Mordo, Moshe Tennenholtz, and Oren Kurland. Sponsored question answering. In *Proceedings of the ACM SIGIR International Conference on the Theory of Information Retrieval*, 2024.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS 2022*, pages 27730–27744, 2022.

[Papadimitriou *et al.*, 2008] Christos Papadimitriou, Michael Schapira, and Yaron Singer. On the hardness of being truthful. In *FOCS 2008*, pages 250–259, 2008.

[Ramesh *et al.*, 2021] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML 2021*, pages 8821–8831, 2021.

[Roughgarden and Talgam-Cohen, 2016] Tim Roughgarden and Inbal Talgam-Cohen. Optimal and robust mechanism design with interdependent values. *ACM Transactions on Economics and Computation*, 4 (3):1–34, 2016.

[Soumalias *et al.*,] Ermis Soumalias, Michael Curry, and Sven Seuken. Truthful aggregation of llms with an application to online advertising. In *Agentic Markets Workshop at ICML 2024*.

[Sun *et al.*, 2019] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV 2019*, pages 7464–7473, 2019.

[Thoppilan *et al.*, 2022] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Agüera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239, 2022.

[Varian, 2007] Hal R. Varian. Position auctions. *International Journal of Industrial Organization*, 25 (6):1163–1178, 2007.

[Vickrey, 1961] William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16(1):8–37, 1961.

[Wei *et al.*, 2021] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2021.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS 2022*, pages 24824–24837, 2022.

[Wei *et al.*, 2023] Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently. *CoRR*, abs/2303.03846, 2023.

[Wu *et al.*, 2023] Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. Large language models are diverse role-players for summarization evaluation. *arXiv preprint arXiv:2303.15078*, 2023.

[Yu *et al.*, 2022] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022.