The Consensus Game: Language Model Generation via Equilibrium Search

Athul Paul Jacob
MIT
apjacob@mit.edu

Yikang Shen
MIT-IBM Lab
yikang.shen@ibm.com

Gabriele Farina & Jacob Andreas
MIT
{gfarina,jda}@mit.edu

Abstract

We introduce a new, a training-free, game-theoretic procedure for language model decoding. Our approach casts language model decoding as a regularized imperfect-information sequential signaling game—which we term the CONSENSUS GAME—in which a GENERATOR seeks to communicate an abstract correctness parameter using natural language sentences to a DISCRIMINATOR. We develop computational procedures for finding approximate equilibria of this game, resulting in a decoding algorithm we call EQUILIBRIUM-RANKING. Applied to a large number of tasks (including reading comprehension, commonsense reasoning, mathematical problem-solving, and dialog), EQUILIBRIUM-RANKING consistently, and sometimes substantially, improves performance over existing LM decoding procedures—on multiple benchmarks, we observe that applying EQUILIBRIUM-RANKING to LLaMA-7B outperforms the much larger LLaMA-65B and PaLM-540B models. These results highlight the promise of game-theoretic tools for addressing fundamental challenges of truthfulness and consistency in LMs.

1 Introduction

Current language models (LMs) perform quite well on some tasks involving generation or verification of factual assertions—including question answering, fact-checking, and even unconditional text generation. But they are far from perfectly reliable, and there is increasing evidence that LMs actually grow more prone to generating false but frequently repeated statements with increasing scale (McKenzie et al., 2023). Further complicating matters, LMs offer multiple affordances for solving factual generation tasks: they may be used both *generatively* (e.g. by asking for the most probable answer to a question) or *discriminatively* (e.g. by presenting a (question, answer) pair and asking whether the answer is acceptable). Within an LM, these two procedures do not always produce consistent results: generative procedures may fail when probability mass is spread across multiple contradicting answers (Wang et al., 2023; Mitchell et al., 2022), while discriminative procedures may fail due to miscalibration (Han et al., 2022; Chen et al., 2022) or subtle dependence on question wording (Jiang et al., 2020). Given these noisy and often-conflicting signals, how should we distill out an LM's best guess at the truth?

This paper presents an approach for reconciling generative and discriminative LM decoding procedures by formulating decoding as a signaling game (Lewis, 2008) that we call the CONSENSUS GAME. At a high level, this game features a GENERATOR agent that must communicate an abstract `correct` or `incorrect` value to a DISCRIMINATOR agent, but may only do so using a set of candidate natural language strings (Fig. 1). Intuitively, an effective *strategy* for this game (i.e. a joint policy) is one in

R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023.

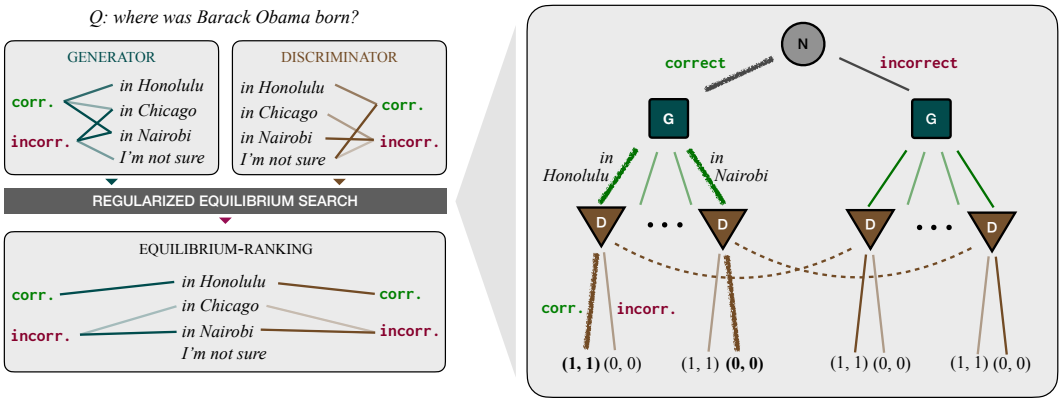

Figure 1: (Left) Example of GENERATOR and DISCRIMINATOR that lacks consensus on the *correctness* of an answer. (Right) Structure of the CONSENSUS GAME, a two-player sequential signaling game with imperfect information. By computing the regularized equilibria of the CONSENSUS GAME, EQUILIBRIUM-RANKING increases consensus between the GENERATOR and DISCRIMINATOR.

which the GENERATOR and DISCRIMINATOR agree on the assignment of strings to correctness values. Given such a strategy, we may inspect it to identify candidates agreed by consensus to be `correct`.

Doing so requires solving a multi-step game with a complex (string-valued) action space. In recent years, *no-regret learning* algorithms have emerged as the preferred technique to compute approximate equilibria in such games, and have been successfully deployed in Poker (Brown & Sandholm, 2018, 2019), Stratego (Perolat et al., 2022), and Diplomacy (Bakhtin et al., 2023; FAIR et al., 2022; Jacob et al., 2022). Here, we show that they can also be applied to free-form language generation tasks. We call this game-theoretic approach to LM decoding EQUILIBRIUM-RANKING. Applied in 6 question answering benchmarks: MMLU (Hendrycks et al., 2020), ARC (Clark et al., 2018), RACE (Lai et al., 2017), HHH (Askell et al., 2021), TruthfulQA (Lin et al., 2021) and, GSM8K (Cobbe et al., 2021), EQUILIBRIUM-RANKING offers substantial improvements over existing generative, discriminative, and mixed decoding procedures. More generally, our results highlight the usefulness of the game-theoretic toolkit for formalizing and improving coherence in LMs. Improved coherence in turn leads to improved accuracy on factual tasks.

2 Language Model Consensus as Equilibrium Search

We study the problem of obtaining correct output from a **language model**, which maps input strings x to output strings y according to some distribution $P_{\text{LM}}(y \mid x)$. While the techniques we present here are general, we focus in this paper on **question answering** problems consisting of a query x (*In which of the following cities was Barack Obama born?*) and a set of candidate answers $\mathcal{Y}$ (*Honolulu, Chicago, . . .*) which may themselves have been sampled from the complete $P_{\text{LM}}(\cdot \mid x)$. Given a set of candidates, we may them with an LM in (at least) two ways:

- *Generatively*, by supplying as input (i) the query x, (ii) the set of candidates $\mathcal{Y}$, and (iii) a natural language prompt indicating that a correct answer is desired. In this case, the LM may be thought of as modeling a distribution $P_{\text{LM}}(y \mid x, \texttt{correct})$, where the token `correct` denotes the fact that the model was prompted to generate a correct answer.
- *Discriminatively*, by supplying as input (i) the query x and (ii) a possible candidate answer $y \in \mathcal{Y}$, together with (iii) a prompt indicating that a correctness assessment $v \in \{\texttt{correct}, \texttt{incorrect}\}$ is sought. In this case, the language model acts as a model of as modeling a distribution $P_{\text{LM}}(v \mid x, y)$ where $v \in \{\texttt{correct}, \texttt{incorrect}\}$.

These two approaches are conceptually equivalent. But as noted in the introduction, current LMs may give very different answers when queried in different ways: answers produced generatively might be assessed `incorrect` with high probability or vice-versa. Research on LMs has proposed two broad solutions to this problem. **Ensembling methods** (Ouyang et al., 2022; Li & Jurafsky, 2016; Glaese et al., 2022) simply combine discriminative and generative scores directly. While moderately

effective, such approaches suffer from the fact that LM predictions are often poorly calibrated both within and across contexts, meaning that scores may not combine in meaningful or consistent ways. **Deliberation methods** (Wei et al., 2022; Yao et al., 2023; Du et al., 2023) perform this reconciliation within the LM itself, e.g. by re-prompting with competing inputs and an instruction to generate a textual justification for the best one. Such methods incur significant computational overhead.[1]

How might we design a principled and computationally efficient procedure for obtaining a "consensus" between competing LM predictions? Informally, a consensus prediction would satisfy two key properties: **coherence** (generative and discriminative scoring procedures should agree about which candidate answers are correct) and **reasonableness** (predictions should not be arbitrary, but instead as close as possible to original generator / discriminator behavior). The key idea in this paper is to operationalize these high-level desiderata in language of game theory, using **regularized equilibrium** concepts as formal framework for defining both coherence and reasonableness. Below, we introduce and explain this framework in detail, describing how to instantiate decoding as a signaling game, then compute equilibrium strategies of this game to obtain consensus LM predictions.

2.1 The CONSENSUS GAME

The CONSENSUS GAME is played on a game tree, as depicted in Figure 1. At the start of the game (that is, at the root of the game tree), a *correctness parameter* $v \in \{\texttt{correct}, \texttt{incorrect}\}$ is selected uniformly at random by the environment. The correctness parameter is observed only by the GENERATOR, and it controls whether the GENERATOR should aim to generate `correct` or `incorrect` answers. Upon observing the correctness parameter, the turn passes to the GENERATOR. The move of the GENERATOR is to then produce a sequence, which is observed by the DISCRIMINATOR, who tries to guess the value of the correctness parameter. The actions available to the DISCRIMINATOR are $\{\texttt{correct}, \texttt{incorrect}\}$, which in our experiments correspond—unless otherwise specified—to the outputs: *"This answer is correct."* and *"This answer is incorrect."* respectively. Payoffs in this game are equal to 1 for both players if the DISCRIMINATOR correctly identifies the value of the correctness parameter, and 0 for both players otherwise.

What makes a good pair of policies for the game? A standard answer to this question in the game theory literature is that a *Nash equilibrium* of the game should be sought. A Nash equilibrium corresponds to a pair of policies—one for the GENERATOR and one for the DISCRIMINATOR—such that each policy is optimal (*i.e.*, it maximizes the corresponding player's expected utility) given the other. Thus, at a Nash equilibrium, no player has an incentive to unilaterally switch to a different policy. Given the way utilities are set in the CONSENSUS GAME, it follows that a Nash equilibrium encodes the first property we are seeking for our method: coherence. However, Nash equilibrium fails the second property, reasonableness. The reason is that the CONSENSUS GAME admits a multitude of Nash equilibria that are incompatible with the common-sense notion of truthfulness. For example, the pair of policies such that the GENERATOR deterministically maps `correct` $\mapsto$ "Nairobi", `incorrect` $\mapsto$ "Honolulu", and the DISCRIMINATOR maps "Nairobi" $\mapsto$ `correct`, "Honolulu" $\mapsto$ `incorrect` form a Nash equilibrium.

In order to sidestep the inappropriate equilibria and ensure reasonableness, we introduce a regularization term in the utility of the players, so that both the GENERATOR and the DISCRIMINATOR are penalized for settling on strategies that are far from their *initial policies*: $\pi_G^{(1)}$ and $\pi_D^{(1)}$. By seeking Nash equilibria of the game with regularized utilities, equilibrium search produces a refinement of the initial GENERATOR and DISCRIMINATOR policies in the direction of increased consensus. In this latter sense, our approach could be described as a *training-free consensus-planning method*.

Similarly to Jacob et al. (2022), we adopt KL divergence from the initial policies of the players as our regularization term in the utility function of the GENERATOR and DISCRIMINATOR. This leads to the following regularized utilities for the CONSENSUS GAME:

$$u_i(\pi_i, \pi_i) \coloneqq -\lambda_i \cdot \mathrm{D}_{\mathrm{KL}}(\pi_i, \pi_i^{(1)}) + \frac{1}{2} \sum_{v \in \{\texttt{correct},\texttt{incorrect}\}} \sum_{y \in \mathcal{Y}} \pi_G(y \mid x, v) \cdot \pi_D(v \mid x, y),$$

for each player $i \in \{\text{GENERATOR}, \text{DISCRIMINATOR}\}$, where λ_i is a coefficient that controls the amount of regularization of the player. We remark that Bakhtin et al. (2023) employed a similar

[1] As shown in Appendix E, they are also orthogonal to, and composable with, the approach we propose here.

regularization method, albeit with the slightly distinct aim of constructing policies compatible with human conventions sampled from a latent distribution, for building an agent that plays no-press Diplomacy. Additionally, it is also worth noting that FAIR et al. (2022) attained human-level performance in full-press Diplomacy by combining language models with a game-solving planner that employs similar regularization techniques, albeit for the purpose of building a game-playing agent. Additional related work is discussed in Appendix C.

2.2 EQUILIBRIUM-RANKING: LM Ranking via equilibrium search

EQUILIBRIUM-RANKING relies on computational game solving techniques—specifically no-regret learning algorithms—to approximate an equilibrium in the CONSENSUS GAME defined above via repeated play. No-regret learning algorithms have emerged in recent years as the preferred technique to approximate equilibria in large games, and have been successfully employed to solve games at human or even superhuman level. No-regret algorithms find equilibrium by repeatedly interacting in the game and refining their policies after each iteration t, so as to minimize regret (i.e. the difference in utility between the chosen action and the action that was optimal in hindisght).

Initial policies At time $t = 1$, that is, before any equilibrium computation has happened, EQUILIBRIUM-RANKING defines the initial policies $\pi_G^{(1)}$ and $\pi_D^{(1)}$ of the GENERATOR and DISCRIMINATOR, respectively, as follows. $\pi_G^{(1)}$ normalizes P_{LM}[2] across v and y: $\pi_G^{(1)}(y \mid x, v) \propto \frac{P_{\text{LM}}(y|x,v)}{\sum_{v'} P_{\text{LM}}(y|x,v')}$. Similarly for the DISCRIMINATOR, the initial policy normalizes across y and v: $\pi_D^{(1)}(v \mid x, y) \propto \frac{P_{\text{LM}}(v|x,y)}{\sum_{y'} P_{\text{LM}}(v|x,y')}$. This crucial step enables us to extract a well calibrated GENERATOR and DISCRIMINATOR from P_{LM}. The specific form of the GENERATOR incorporates $v = \texttt{incorrect}$, and this resembles contrastive decoding (Li et al., 2022), where they rely on a weaker LM as opposed to an LM conditioned on `incorrect` (See, Section 3 for details). This DISCRIMINATOR resembles approaches that query the LM itself to produce critiques (Ganguli et al., 2023; Chen et al., 2023b; Yao et al., 2023). However, to the best of our knowledge, this specific instantiation has not been explored in the past.

Evolution of policies A classic observation in the theory of imperfect-information sequential games is that minimization of regret (viewed as a function of their overall policy on the game tree) can be achieved by solving separate, *local*, regret minimization problems at each information set (*i.e.*, decision point) of the game. This observation underpins the CFR framework (Zinkevich et al., 2007), as well as its generalization to more general convex losses, known as laminar regret decomposition (Farina et al., 2019). In our case, these techniques enable us to decompose the policy update of the players into separate updates for each correctness parameter v (for the GENERATOR) and for each sequence y (for the DISCRIMINATOR). We provide more detail and background in Appendix A.

In our setting, after operating the regret decomposition step, we find that the local regret minimization problems are composed of a bilinear term, plus a strongly convex KL-regularization term. Such composite utilities can be handled by the piKL algorithm (Jacob et al., 2022), which is specifically designed to perform regret minimization on KL-regularized objectives. In our setting, piKL prescribes that each player keep track of their average values

$$Q_G^{(t)}(y \mid x, v) \coloneqq \frac{1}{2t} \sum_{\tau=1}^{t} \pi_D^{(\tau)}(v \mid x, y), \qquad Q_D^{(t)}(v \mid x, y) \coloneqq \frac{1}{2t} \sum_{\tau=1}^{t} \pi_G^{(\tau)}(y \mid x, v).$$

Each player then updates their policy according to the rules

$$\pi_G^{(t+1)}(y \mid x, v) \propto \exp\left\{ \frac{Q_G^{(t)}(y \mid x, v) + \lambda_G \log \pi_G^{(1)}(y \mid x, v)}{1/(\eta_G t) + \lambda_G} \right\}, \tag{1}$$

$$\pi_D^{(t+1)}(v \mid x, y) \propto \exp\left\{ \frac{Q_D^{(t)}(v \mid x, y) + \lambda_D \log \pi_D^{(1)}(v \mid x, y)}{1/(\eta_D t) + \lambda_D} \right\}, \tag{2}$$

[2]In ARC, RACE, HHH, TruthfulQA, and GSM8K, based on prior work (Touvron et al., 2023; Brown et al., 2020), we additionally normalize $P_{\text{LM}}(u|x)$ by the likelihood of the completion given "Answer:" as context: $P_{\text{LM}}(u \mid$ "Answer:").

where $\eta_G, \eta_D > 0$ are *learning rate* hyperparameters. piKL no-regret dynamics are known to have strong guarantees, including the following (more formal statements are available in Appendix A):

- Regularization toward reasonableness. The average policy of any player remains within a radius of size roughly $1/\lambda_i$ from the initial policy $\pi_i^{(1)}$, where λ_i is the amount of regularization applied to any player $i \in \{\text{GENERATOR}, \text{DISCRIMINATOR}\}$ (see Proposition 3).
- Convergence to consensus. The average correlated distribution of play of GENERATOR and DISCRIMINATOR converges to the set of (regularized) coarse-correlated equilibria of the game.

At convergence, EQUILIBRIUM-RANKING returns π_G^* and π_D^*, which are the refined GENERATOR and DISCRIMINATOR. We also remark that a recent result by Anagnostides et al. (2022) showed that as long as the regularization function has Lipschitz-continuous gradients (a condition that can be easily achieved by introducing a small perturbation in the KL regularization term), dynamics such as piKL converge in iterates to a regularized Nash equilibrium of the game. In practice, we do not investigate introducing a small perturbation, as we witness good convergence properties in practice even without any perturbation. As mentioned earlier, convergence to a regularized Nash equilibrium is important to guarantee both coherence and reasonableness. Extensive empirical validation presented in the next section demonstrates the benefits of this approach in practice.

3 Experiments

As discussed in the previous section, EQUILIBRIUM-RANKING focuses on improving the *correctness* of language models in question-answering (QA) tasks. However, correctness manifests in various forms across different domains, including truthfulness, factuality, valid reasoning, value alignment, among others. Therefore, we will evaluate its performance on a diverse set of QA tasks: MMLU (Hendrycks et al., 2020), ARC (Clark et al., 2018), RACE (Lai et al., 2017), HHH (Askell et al., 2021), and TruthfulQA (Lin et al., 2021). It's important to note that EQUILIBRIUM-RANKING is a sampling strategy and not a delibration method like chain-of-thought (CoT) (Wei et al., 2022) and self-consistency (Wang et al., 2023). Nevertheless, we will demonstrate in GSM8K (Cobbe et al., 2021) that we can achieve some additional gains when combining EQUILIBRIUM-RANKING with self-consistency and CoT. The results on TruthfulQA and GSM8K are presented in Appendix E.

Hyperparameters EQUILIBRIUM-RANKING has four parameters, η_D, λ_D and η_G, λ_G. Although tuning these parameters will lead to better performance, in all our experiments we set $\eta_D = \lambda_D = \eta_G = \lambda_G = 0.1$. Furthermore, we run EQUILIBRIUM-RANKING for 5000 iterations.

Actions in the CONSENSUS GAME As mentioned in Section 2, in order to make our approach amenable to current computational techniques, we make the modeling assumption that the GENERATOR picks distribution over a finite set of candidates $\mathcal{Y}$. In multiple-choices tasks, these are the multiple choice options. In generative tasks, a common approach to generate the finite set of candidates is via sampling with nucleus (Holtzman et al., 2019) and top-k (Fan et al., 2018b) from the distribution $P_{\text{LM}}(y \mid q, \texttt{correct})$ where $y \in \mathcal{Y}$. This is exactly the approach we use in our experiments, with $p = 0.9$ for nucleus sampling and $k = 50$.

Models We use the 7B and 13B parameter models from the LLaMA family (Touvron et al., 2023) and perform 16-bit inference for all our experiments.

Prompting for `correct` and `incorrect` answers In our work, unless otherwise specified, conditioning on $(x, \texttt{correct})$ for the P_{LM} corresponds to the standard zero-shot prompt. Similarly, conditioning on $(x, \texttt{incorrect})$ is similar to $(x, \texttt{correct})$ with the only difference that *"Answer:"* is replaced with *"Incorrect Answer:"* in the prompt.

Baselines In the multiple-choice based datasets (ARC, RACE, HHH, MMLU), we consider the following approaches: **Generative Ranking (G):** This baseline (Brown et al., 2020; Touvron et al., 2023) ranks every candidate y by $P_{\text{LM}}(y \mid x, \texttt{correct})$ and picks the top candidate. This is the standard approach used in past work. Due to implementational differences, when available, we include both official scores and our version. **Mutual Information Ranking (MI):** The mutual information based (Li & Jurafsky, 2016) baseline is an ensemble-based approach that reweights every candidate y by $P_{\text{LM}}(y \mid x, \texttt{correct}) \cdot P_{\text{LM}}(\texttt{correct} \mid x, y)$. **Self-Contrastive Ranking (SC):** This approach utilizes

Domain	Model	G*	G	MI	SC	D	Equil. ranking ER-G	Equil. ranking ER-D
MMLU	LLaMA-7B	-	30.4	33.1	30.5	**40.4**	39.4	39.9
	LLaMA-13B	-	41.7	41.8	41.7	41.9	44.9	**45.1**
ARC Easy	LLaMA-7B	72.8	68.2	68.8	69.5	52.5	**71.6**	71.5
	LLaMA-13B	74.8	71.2	71.5	73.0	65.0	76.1	**76.4**
ARC Challenge	LLaMA-7B	47.6	47.3	47.4	56.5	42.7	**58.7**	58.3
	LLaMA-13B	52.7	51.9	52.1	59.3	48.5	61.1	**61.4**
RACE Middle	LLaMA-7B	61.1	57.7	57.7	60.4	51.5	63.2	**63.5**
	LLaMA-13B	61.6	60.1	60.2	64.8	58.3	67.9	**68.6**
RACE High	LLaMA-7B	46.9	46.4	46.3	53.1	46.0	56.3	**56.4**
	LLaMA-13B	47.2	47.9	48.4	58.9	55.1	62.1	**62.8**
HHH	LLaMA-7B	-	59.3	57.9	67.4	70.1	**71.5**	**71.5**
	LLaMA-13B	-	60.2	59.7	57.9	**69.2**	61.1	61.1

Table 1: Results of the different approaches across multiple tasks. We compute the accuracies on the test set of these benchmarks. EQUILIBRIUM-RANKING outperforms other approaches on most tasks. EQUILIBRIUM-RANKING performs well, even in cases where one of GENERATOR or DISCRIMINATOR is far worse than the other. * indicates the results from Touvron et al. (2023)

the normalized generator $\pi_G^{(1)}$ to reweight every candidate y by $\pi_G^{(1)}(\texttt{correct} \mid x, y)$. As discussed in Section 2, this shares similarities with contrastive decoding (Li et al., 2022). **Discriminative Ranking (D):** This approach reweights every query-candidate pair (x, y) by $\pi_D^{(1)}(\texttt{correct} \mid x, y)$. **Equilibrium Ranking Generator (ER-G):** Similar to SC, this approach utilizes the final EQUILIBRIUM-RANKING-based generator π_G^* to reweight every candidate y by $\pi_G^*(y \mid x, \texttt{correct})$. **Equilibrium Ranking Discriminator (ER-D):** Similar to D, this approach utilizes the final EQUILIBRIUM-RANKING-based discriminator π_D^*. This approach reweights every query-candidate pair (x, y) by $\pi_D^*(\texttt{correct} \mid x, y)$. In free-form text generation tasks (TruthfulQA, GSM8K), we additionally consider **greedy decoding**. In the mathematical reasoning task (GSM8K), we also consider **self-consistency** (Wang et al., 2023).

The application of EQUILIBRIUM-RANKING-based approaches consistently yields improved results, surpassing or at least matching the performance of all baseline approaches across various tasks. This robustness is particularly interesting, as it demonstrates that EQUILIBRIUM-RANKING is adept at handling diverse scenarios, even in situations when the initial GENERATOR or DISCRIMINATOR are not effective. As EQUILIBRIUM-RANKING is a sampling strategy, it can even be combined with deliberation methods like self-consistency (Wang et al., 2023) or tree-of-thought (Yao et al., 2023). Finally, we note that EQUILIBRIUM-RANKING demonstrates computational efficiency by eliminating the need for repetitive queries to language models. As noted in Appendix D, we highlight that across multiple tasks (ARC, RACE), LLaMA-13B and LLaMA-7B with EQUILIBRIUM-RANKING is even able to outperform larger models: LLaMA-65B and PaLM-540B. Further details may be found in Appendix D and Appendix E.

4 Conclusion

This paper presents an approach for reconciling generative and discriminative LM decoding procedures by formulating decoding as an imperfect-information signaling game between a GENERATOR and a DISCRIMINATOR, called the CONSENSUS GAME. EQUILIBRIUM-RANKING leverages computational game solving techniques to compute an approximate equilibria of this game. When applied to 6 diverse question answering benchmarks: MMLU, ARC, RACE, HHH, TruthfulQA and, GSM8K, we note that EQUILIBRIUM-RANKING offers substantial improvements over existing generative, discriminative, and mixed decoding procedures. In particular, we also observe that

applying EQUILIBRIUM-RANKING to LLaMA-7B can outperform much larger LLaMA-65B and PaLM-540B models

References

Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. On last-iterate convergence beyond zero-sum games. In *International Conference on Machine Learning*, 2022.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Anton Bakhtin, David J Wu, Adam Lerer, Jonathan Gray, Athul Paul Jacob, Gabriele Farina, Alexander H Miller, and Noam Brown. Mastering the game of no-press Diplomacy via human-regularized reinforcement learning and planning. In *The Eleventh International Conference on Learning Representations*, 2023.

Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456): 885–890, 2019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*, 2023a.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*, 2023b.

Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. A close look into the calibration of pre-trained language models. *arXiv preprint arXiv:2211.00151*, 2022.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A Saurous, Jascha Sohl-Dickstein, et al. Language model cascades. *arXiv preprint arXiv:2207.10342*, 2022.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Meta FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018a. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, 2018b.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Online convex optimization for sequential decision processes and extensive-form games. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, Karina Nguyen, Liane Lovitt, Michael Sellitto, Nelson Elhage, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robert Lasenby, Robin Larson, Sam Ringer, Sandipan Kundu, Saurav Kadavath, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Christopher Olah, Jack Clark, Samuel R. Bowman, and Jared Kaplan. The capacity for moral self-correction in large language models, 2023.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

Zhixiong Han, Yaru Hao, Li Dong, Yutao Sun, and Furu Wei. Prototypical calibration for few-shot learning of language models. In *The Eleventh International Conference on Learning Representations*, 2022.

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

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.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rygGQyrFvH`.

Athul Paul Jacob, David J. Wu, Gabriele Farina, Adam Lerer, Hengyuan Hu, Anton Bakhtin, Jacob Andreas, and Noam Brown. Modeling strong and human-like gameplay with kl-regularized search. In *International Conference on Machine Learning*, 2022.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*, 2017.

David Lewis. *Convention: A philosophical study*. John Wiley & Sons, 2008.

Jiwei Li and Dan Jurafsky. Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372*, 2016.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*, 2022.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023.

Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, et al. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*, 2023.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling, 2023.

Eric Mitchell, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher D Manning. Enhancing self-consistency and performance of pre-trained language models through natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1754–1768, 2022.

Long Ouyang, Jeffrey 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T Connor, Neil Burch, Thomas Anthony, et al. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

Jianhao Shen, Yichun Yin, Lin Li, Lifeng Shang, Xin Jiang, Ming Zhang, and Qun Liu. Generate & rank: A multi-task framework for math word problems. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2269–2279, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.195. URL https://aclanthology.org/2021.findings-emnlp.195.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

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, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen 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 Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. Lamda: Language models for dialog applications, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

Martin Zinkevich, Michael Bowling, Michael Johanson, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Neural Information Processing Systems (NIPS)*, 2007.

A Details about Regret and Regret Decomposition Methods

After each repetition t of the game, each player—in this case, the GENERATOR and the DISCRIMINATOR—refines their policies, in such a way that throughout the course of time, the *regrets*

$$\mathrm{Reg}_G^{(T)} \coloneqq \max_{\pi_G^*} \left\{ \sum_{t=1}^{T} u_G(\pi_G^*, \pi_D^{(t)}) - \sum_{t=1}^{T} u_G(\pi_G^{(t)}, \pi_D^{(t)}) \right\}, \tag{3}$$

$$\mathrm{Reg}_D^{(T)} \coloneqq \max_{\pi_D^*} \left\{ \sum_{t=1}^{T} u_D(\pi_G^{(t)}, \pi_D^*) - \sum_{t=1}^{T} u_D(\pi_G^{(t)}, \pi_D^{(t)}) \right\}, \tag{4}$$

cumulated by the players are guaranteed to grow sublinearly as a function of the number of rounds of learning T.

As mentioned in the body, a classic observation in the theory of imperfect-information sequential games is that minimization of regret (viewed as a function of their overall policy on the game tree) can be achieved by solving separate, *local*, regret minimization problems at each information set (*i.e.*, decision point) of the game. In our case, these techniques enable us to decompose the policy update of the players into separate updates for each correctness parameter v (for the GENERATOR) and for each sequence y (for the DISCRIMINATOR). More specifically, suppose that the GENERATOR updates their policies $\pi_G^{(t)}(\cdot \mid x, v)$ independently for each correctness parameter $v \in \{\texttt{correct}, \texttt{incorrect}\}$ they might receive, seeking to independently minimize regret

$$\mathrm{Reg}_G^{(T)}(v) \coloneqq \max_{\pi^* \in \Delta(\mathcal{Y})} \left\{ \sum_{t=1}^{T} \tilde{u}_G^{(t)}(\pi^* \mid x, v) - \tilde{u}_G^{(t)}(\pi_G^{(t)}(\cdot \mid x, v) \mid x, v) \right\}$$

with respect to the following *counterfactual utility functions*

$$\tilde{u}_G^{(t)}(\pi_G \mid x, v) \coloneqq -\lambda_G \mathrm{D}_{\mathrm{KL}} \left(\pi_G(\cdot \mid x, v) \,\middle\|\, \pi_G^{(0)}(\cdot \mid x, v) \right) + \frac{1}{2} \sum_{y \in \mathcal{Y}} \pi_D^{(t)}(v \mid x, y) \cdot \pi_G(y \mid x, v) \tag{5}$$

for all v. Then, it is known that when these independent goals are met for all v, so is the goal of keeping regret (3) subliner, and in particular

$$\mathrm{Reg}_G^{(T)} \leq \mathrm{Reg}_G^{(T)}(\texttt{correct}) + \mathrm{Reg}_G^{(T)}(\texttt{incorrect})$$

no matter the time horizon T. Similarly, when the DISCRIMINATOR seeks to update their policy $\pi_D^{(t)}(\cdot \mid x, y)$ for each $y \in \mathcal{Y}$ independently, so as to minimize regret

$$\mathrm{Reg}_D^{(T)}(y) \coloneqq \max_{\pi^* \in \Delta(\{\texttt{correct},\texttt{incorrect}\})} \left\{ \sum_{t=1}^{T} \tilde{u}_D^{(t)}(\pi^* \mid x, y) - \tilde{u}_D^{(t)}(\pi_D^{(t)}(\cdot \mid x, y) \mid x, y) \right\}$$

with respect to the counterfactual utility functions

$$\tilde{u}_D^{(t)}(\pi_D \mid x, y) \coloneqq -\lambda_D \mathrm{D}_{\mathrm{KL}} \left(\pi_D(\cdot \mid x, y) \,\middle\|\, \pi_D^{(0)}(\cdot \mid x, y) \right) + \frac{1}{2} \sum_{v \in \{\texttt{correct}, \texttt{incorrect}\}} \pi_G^{(t)}(v \mid x, y) \cdot \pi_D(v \mid x, y),$$

then their overall regret $\mathrm{Reg}_D^{(T)}$ satisfies

$$\mathrm{Reg}_D^{(T)} \leq \sum_{y \in \mathcal{Y}} \mathrm{Reg}_D^{(t)}(y).$$

The counterfactual utilities $\tilde{u}_G$ and $\tilde{u}_D$ defined above are composed of a bilinear term and a strongly convex KL-regularization term. To guarantee sublinear regret with respect to such utility functions, we use the piKL algorithm Jacob et al. (2022).

A.1 Guarantees of the piKL No-Regret Dynamics

We recall the following known properties of the piKL algorithm. An informal description of the guarantees was included in Section 2.

Proposition 1 (Bakhtin et al. 2023)**.** *Let $i \in \{\text{GENERATOR}, \text{DISCRIMINATOR}\}$ be any player. After any T iterations of training, the regret $\text{Reg}_i^{(T)}$ cumulated by the policies $\pi_i^{(t)}$ of player i produced by piKL, is upper bounded by only a logarithmic quantity in T. More precisely, $\text{Reg}_i^{(T)} = O\left(\min\left\{\frac{2\log T}{\lambda_i}, T\eta_i\right\} + \frac{\log|\mathcal{Y}|}{\eta_i}\right)$, where the asymptotic O notation hides constants independent on the time horizon T, learning rate η_i of the player, and regularization coefficient λ_i.*

Proposition 2 (Folklore connection between regret minimization and equilibria)**.** *The empirical frequency of play converges to the set of coarse correlated equilibria of the* CONSENSUS GAME*.*

Proposition 3 (Jacob et al. 2022)**.** *The average policy $\bar{\pi}_i^{(T)}$ of player $i \in \{\text{GENERATOR}, \text{DISCRIMINATOR}\}$ produced by piKL after T iterations is guaranteed to be within a radius proportional to λ_i centered in the initial policy $\pi_i^{(1)}$. More precisely, $\text{D}_{\text{KL}}(\bar{\pi}_i^{(T)}, \pi_i^{(1)}) \leq \frac{1}{\lambda_i}(1 + o(1))$, where the asymptotic notation $o(1)$ denotes a quantity decreasing, as a function of the time horizon T, at a rate of $\log T/(\lambda_i T)$.*

B HHH Prompts

In the HHH experiments, $(x, \texttt{correct})$ corresponds to the prompt:

```
"You are a helpful, honest and harmless assistant.
Human: {x}
Assistant:"
```

And $(x, \texttt{incorrect})$ corresponds to the prompt:

```
"You are a unhelpful, dishonest and harmful assistant.
Human: {x}
Assistant:"
```

C Other Related Work

Many decoding strategies have been proposed for language models, such as top-k sampling (Fan et al., 2018a), nucleus sampling (Holtzman et al., 2020), and typical sampling (Meister et al., 2023). These methods primarily focus on producing diverse, high-quality text from a language model. However, they decode from the LM without any emphasis on the correctness of the generated sequences. As we show in Section 3, EQUILIBRIUM-RANKING is naturally complementary and be combined with any of these sampling strategies.

Re-ranking is a common approach for selecting the correct answer from a set of candidates sampled from LM. Cobbe et al. (2021) train a verifier to re-ranked the sampled outputs. Shen et al. (2021) jointly train a ranking model with the generation model to improve the model accuracy. Thoppilan et al. (2022) collect additional human annotations to train the ranking model for response filtering. As we discuss in Section 2, our work focuses on leveraging an existing LM and using them in a training-free manner as a discriminator. However, we note that we do not make any specific assumption on the specific form of a the GENERATOR or DISCRIMINATOR. As such, EQUILIBRIUM-RANKING can be combined with these approaches.

As previously mentioned, EQUILIBRIUM-RANKING differs from recent deliberation methods, as highlighted in various recent work (Wei et al., 2022; Madaan et al., 2023; Shinn et al., 2023; Yao et al., 2023; Dohan et al., 2022). In Section 3, we demonstrate how EQUILIBRIUM-RANKING can be integrated with these approaches. In another line of work, Du et al. (2023); Chen et al. (2023a) employ multiple instances of language models suggest and "debate" individual responses and reasoning processes across multiple iterations, ultimately converging on a shared final answer. In contrast, EQUILIBRIUM-RANKING can be viewed as a variant of this multi-agent debate, wherein the "debate" occurs within the regret-minimization framework rather than in the context of language models.

D Results Discussion

MMLU The massive multi-task language understanding benchmark (MMLU) (Hendrycks et al., 2020) is used to measure language model's multitask accuracy. It consists of questions in the multiple choice format across a wide variety of subdomains in social sciences, humanities and STEM. We evaluate our models in the zero-shot setting following the format described in Hendrycks et al. (2020); Touvron et al. (2023) and the results are presented in the first row of Table 1. For both LLaMA-7B and LLaMA-13B, the EQUILIBRIUM-RANKING-based approaches matches or outperforms all other baselines. In fact, zero-shot LLaMA-7B with ER-D (39.9) outperforms 5-shot LLaMA-7B (35.1), while zero-shot LLaMA-13B with ER-D (45.1) is competitive with 5-shot LLaMA-13B (46.9). LLaMA-7B with ER-D (39.9) even outperforms zero-shot GPT3-175B (37.7) (Hendrycks et al., 2020), while zero-shot LLaMA-13B with ER-D (45.1) outperforms 5-shot GPT3-175B (43.9) (Hendrycks et al., 2020).

ARC The AI2 Reasoning Challenge (ARC) (Clark et al., 2018) is an advanced question answering dataset used to study a model's knowledge and reasoning abilities based on grade school science questions. It is split in to two subcategories: easy (ARC-Easy) and challenge (ARC-Challenge). The challenge set was constructed as the set of questions that were answered incorrectly by retrieval and word co-occurence based algorithms. The results are presented in second and third rows of Table 1. On ARC-Easy, ER-D outperforms our implementation of generative ranking. We also note that LLaMA-13B with ER-D (76.4) outperform all the baseline approaches and is even competitive with the much larger PaLM-540B model (76.6) Chowdhery et al. (2022). On ARC-Challenge, ER-D performs quite well, significantly outperforming all the baseline approaches. We also note that LLaMA-7B with ER-D (58.3) and LLaMA-13B with ER-D (61.4) outperforms even the much larger models: LLaMA-65B (56.0) Touvron et al. (2023) and PaLM-540B (53.0)(Chowdhery et al., 2022) by upto 8%.

RACE RACE is a reading comprehension benchmark introduced in Lai et al. (2017) collected from English examinations of middle and high school students. The dataset is correspondingly split into RACE-middle and Race-high. The dataset consists of a passage followed by questions. The passages were constructed for evaluating student's English reasoning and understanding ability. The results on this benchmark is presented in rows 4 and 5 of Table 1. On RACE-middle, like before, ER-D based models outperforms all the baselines. We note that LLaMA-13B with ER-D (68.6) even outperforms much larger models: LLaMA-65B (67.9) Touvron et al. (2023) and PaLM-540B (68.1)(Chowdhery et al., 2022). On RACE-high, we have a similar story as with ARC-C. ER-D outperforms all baselines. LLaMA-7B with ER-D (56.4) is able to significantly outperform much larger models: LLaMA-65B (51.6) Touvron et al. (2023) and PaLM-540B (49.1) (Chowdhery et al., 2022).

HHH HHH (Helpful, Honest and Harmless) (Srivastava et al., 2023; Askell et al., 2021) is a dataset of 200 multiple-choice designed to measure LM alignment with high-level quality guidelines. We use a different set of prompts for this task for the GENERATOR. These are defined in Appendix B. The results on this benchmark is presented in the last row of Table 1. LLaMA-7B with ER-D outperforms the baseline methods. And although LLaMA-13B with ER-D with the default parameter performs worse than discriminative ranking (D) (69.2), we note ER-D with $\lambda_G = 0.01$ and $\lambda_D = 1.0$ achieves an accuracy of 70.6%.

E Additional Results: TruthfulQA, GSM8K

TruthfulQA TruthfulQA (Lin et al., 2021) is a benchmark consisting of over 800 questions across a multitude of domains that were crafted to encourage humans to answer them incorrectly due to misconceptions. The dataset evaluates a model's ability to not generate false answers learnt from imitation learning on text. On this task, we consider **greedy decoding** in addition to our other ranking-based approaches. In this setting, 10 candidate sequences are sampled using nucleus and top-k sampling. These candidates are then ranked based on the approaches we described earlier. The results on the test set are presented in Table 2. Based on past work (Lin et al., 2021), we measure BLEU accuracy (BLEU-Acc). For a sequence a, the BLEU-Acc over reference correct candidates b_{correct} and reference incorrect candidates b_{correct} is computed as follows:

$$\text{BLEU-Acc}(a) := \mathbb{I}(\text{BLEU}(a, b_{\text{correct}}) > \text{BLEU}(a, b_{\text{incorrect}})) \tag{6}$$

Where, BLEU(a, b) computes the BLEU (Papineni et al., 2002) score of a candidate string a over a set of reference candidates b. With LLaMA-7B, we observe only modest improvements for ER-G and ER-D over the greedy baseline. However, with LLaMA-13B, we note increased scores for both methods. This benchmark is known to exhibit negative scaling Lin et al. (2021) (performance drop as the model size increases). As evidenced by the performance difference with ER-G between LLaMA-7B and LLaMA-13B, we note that EQUILIBRIUM-RANKING is in fact capable of countering this behaviour.

Domain	Model	Greedy	MI	SC	D	**Equil. ranking** ER-G	ER-D
TruthfulQA	LLaMA-7B	33.41	34.79 $\pm$ 0.90	**34.91** $\pm$ 0.57	34.17 $\pm$ 1.19	34.61 $\pm$ 0.99	34.27 $\pm$ 0.39
	LLaMA-13B	33.05	36.30 $\pm$ 0.37	34.61 $\pm$ 1.33	39.05 $\pm$ 1.42	**39.83** $\pm$ 2.20	38.63 $\pm$ 1.76

Table 2: Results on TruthfulQA (Generative). Average BLEU-Acc results on the held-out set across 5 runs. LLaMA-13B with ER-G outperforms or is on par with all baselines. **MI**: Mutual Information Ranking, **SC**: Self-Contrastive Ranking, **D**: Discriminative Ranking, **ER-G**: Equilibrium Ranking Generator, **ER-D**: Equilibrium Ranking Discriminator. $\pm$ indicates 1 standard deviation computed across 5 runs.

GSM8K In our last set of experiments, we consider grade-school math (GSM8K) Cobbe et al. (2021), a popular benchmark used to study model's mathematical reasoning ability. We use this benchmark to study whether we can combine our approach with chain-of-thought (Wei et al., 2022). As we described earlier, we generate 20 candidate reasoning paths sampled using nucleus and top-k using the 8-shot setup proposed in Wei et al. (2022). We employ self-consistency (Wang et al., 2023) over the candidate sequences, where we score each reasoning path with our baselines. Finally, we aggregate the scores for each answer corresponding to the reasoning paths and pick the answer with the highest score. In Table 3, we present the results. We note that EQUILIBRIUM-RANKING-based approaches performs on par or slightly better compared to self-consistency (majority vote).

Domain	Model	Greedy	MV	MI	SC	D	**Equil. ranking** ER-G	ER-D
GSM8K	LLaMA-7B	10.8	14.7 $\pm$ 0.2	14.6 $\pm$ 0.5	13.4 $\pm$ 0.3	15.0 $\pm$ 0.6	13.0 $\pm$ 0.5	**15.1** $\pm$ 0.6
	LLaMA-13B	14.9	22.5 $\pm$ 0.5	22.5 $\pm$ 0.8	**23.1** $\pm$ 0.5	22.5 $\pm$ 0.6	22.5 $\pm$ 0.6	23.0 $\pm$ 0.5

Table 3: Average accuracy of methods on the test set of GSM8K acroos 5 runs. In all cases, except greedy, 20 candidates were sampled. EQUILIBRIUM-RANKING-based approaches performs on par or slightly better compared to the majority vote baseline. **MV**: Majority Vote, **MI**: Mutual Information Ranking, **SC**: Self-Contrastive Ranking, **D**: Discriminative Ranking, **ER-G**: Equilibrium Ranking Generator, **ER-D**: Equilibrium Ranking Discriminator. $\pm$ indicates 1 standard deviation.