From Competition to Coordination: Market Making as a Scalable Framework for Safe and Aligned Multi-Agent LLM Systems

Brendan Gho, Suman Muppavarapu, Afnan Shaik, Tyson Tsay, Atharva Mohan
James Begin, Vasu Sharma, Kevin Zhu

Algoverse AI Research

Abstract

Large Language Models (LLMs) are beginning to collaborate, debate, and negotiate with one another, opening new paths toward collective intelligence, but also new risks in coordination, alignment, and truthfulness. In this paper, we present how market mechanisms can structure these multi agent interactions in a way that makes truthful reasoning an emergent property rather than a hand crafted rule.

We propose a market making framework where each LLM agent acts as a market maker or trader, continuously updating and exchanging probabilistic beliefs through negotiation. Instead of enforcing agreement from the top down, the system self-organizes: agents are rewarded for offering accurate and consistent information, and penalized when their beliefs fail to hold up under scrutiny. The result is a decentralized process where truth emerges from incentive alignment, not from central oversight.

Through experiments across factual reasoning, estimation, and multi-step analytical tasks, we find that market based coordination consistently improves collective truthfulness and reasoning accuracy, often by more than 10% compared to traditional debate or majority vote frameworks. Beyond empirical gains, our findings suggest that economic principles like liquidity, price discovery, and arbitrage can serve as powerful design tools for building safer, more transparent, and more self correcting LLM societies.

1 Introduction

The rapid deployment of artificial intelligence systems across safety-critical domains has intensified concerns regarding existential risks, particularly the emergence of deceptively aligned models (Hendrycks, Mazeika, and Woodside 2023). Recent evidence demonstrates that advanced language models exhibit strategic deception, including attempts to game evaluation protocols and misrepresent their internal states during training (Carlsmith 2023). These alignment failures manifest as sycophancy, systematic untruthfulness, and adversarial behaviour, actions that empirically worsen with increased model scale (Ji et al. 2025).

Existing alignment methodologies face fundamental limitations. Reinforcement Learning from Human Feedback (RLHF), while effective for surface-level behavioural modification, remains vulnerable to reward haking and evaluator deception. Debate-based approaches require human adjudication that cannot scale to superhuman reasoning capabilities. The Alignment Research Center's Eliciting Latent Knowledge (ELK) framework defines this challenge: extracting model's true internal representations rather than their strategically chosen outputs (Christiano, Xu et al. 2022).

This paper explores market making as a novel method for alignment and truth elicitation. Inspired by economic prediction markets, the approach involves a market maker that continuously offers prices on propositions and traders that buy or sell based on their beliefs. Through a process of iterative trading, prices converge to a probability that reflects the collective belief about ground truth. Models take the role of traders, updating the "market probability" as they present new evidence or reasoning steps. This operation of trading incentivizes truthful contributions in order to receive the most profitable trade, improving the market's accuracy. The framework also allows for myopic trader agents who are blind to past information, preventing long-term scheming and manipulation. By converting truth-seeking into an equilibrium of incentives rather than a contest of persuasion or subjective judgment, market making offers a potentially robust and scalable alternative to debate and oversight for eliciting honest beliefs from advanced AI systems.

2 Related Works

2.1 The Challenges of Control

(Amodei et al. 2016) established a taxonomy of AI safety failure modes comprising five critical categories: negative side effects, reward hacking, scalable oversight limitations, unsafe exploration, and distributional shift vulnerabilities. This foundational framework reveals an inherent tension in alignment objectives: excessive optimization for harmlessness produces ineffectual systems, while prioritizing capability enables potential misuse (Bai et al. 2022). The multidimensional nature of these constraints implies that no single methodology can simultaneously address all failure modes, necessitating approaches that optimize across multiple safety dimensions.

2.2 Human-Centric Alignment

Early alignment techniques relied on direct human supervision through iterative feedback mechanisms. Reinforcement Learning from Human Feedback (RLHF) exemplifies this paradigm, wherein human evaluators shape model behaviour through preference rankings (Bai et al. 2022). AI Safety via Debate was proposed by (Irving, Christiano, and Amodei 2018), structuring oversight as adversarial argumentation adjudicated by human judges. These approaches face three fundamental limitations. First, the bandwidth constraint: human evaluation cannot scale to the volume and velocity of decisions required in deployed systems (Amodei et al. 2016). Second, the competence boundary: superhuman AI capabilities exceed human evaluators' ability to assess correctness (Ji et al. 2025). Third, the alignment targeting problem, where models optimize for evaluator approval rather than ground truth, leading to sycophantic behaviour and strategic deception (Carlsmith 2023; Park et al. 2023).

2.3 AI-Mediated Oversight

These challenges of human-centric alignment motivated AI-mediated oversight using secondary AI systems to help supervise the model being tested, aiming to augment or replace human adjudicators. JudgeLM replaces the human judge in AI debate with an AI system, showing extended capabilities in a variety of situations (Zhu, Wang, and Wang 2025). Bowman et al. also explore how less-capable AIs can reliably evaluate stronger ones without expert human intervention (Bowman et al. 2022). Together, these efforts suggest that scalable, AI-mediated oversight may be a necessary step toward maintaining safe and reliable control as AI capabilities continue to grow.

2.4 Market Making as a Control Mechanism

Market making offers an incentive-based alternative to adjudication. In this method, an automated market maker posts prices for propositions and traders (models or submodels) buy or sell claims based on their beliefs; iterative trading drives prices toward an equilibrium that reflects collective credence (Holmes 2020). A key intended advantage is enforcing myopic behavior where trader agents optimize per-step trades, reducing incentives for long-term scheming that can undermine debate or RLHF. Market mechanisms also facilitate per-step inspection, probabilistic scoring, and potential scalability without continuous human adjudication. Thus, market making attempts to improve upon existing methods of AI control.

Practical challenges to market making do remain, including defending the market maker against false claims, designing proper rewards that prevent gaming, and ensuring robust performance when truth is hazy. Existing work is primarily a toy implementation of market making from Cameron Holmes (Holmes 2020).

3 Methodology

We implement market making using two agents: a market-maker model, $\mathbf{M}$, and a trader model of the same model. Market making begins with $\mathbf{M}$ providing an initial judgment consisting of:

1. a claim,
2. supporting reasoning, and
3. a prediction value $p_0 \in [0, 1]$ quantifying the claim.

Given $\mathbf{M}$'s judgment, the trader model then generates an argument intended to maximally shift $\mathbf{M}$'s prediction value. This is analogous to a trader introducing new information to change the market price.

Each subsequent iteration proceeds with $\mathbf{M}$ producing a new judgment while also considering the trader's previous arguments. Exact prompting details are provided in **Appendix A**. The cycle repeats until the market maker has provided at most $\mathbf{N}$ judgments or has reached an equilibrium; we consider an equilibrium to have been reached when the range of the last three prediction values satisfies

$$\max\{p_{t-2}, p_{t-1}, p_t\} - \min\{p_{t-2}, p_{t-1}, p_t\} \leq T,$$

where T is a threshold constant. In our experiments, we set $N = 10$ and $T = 0.2$. Finally, we measure the impact of market making by comparing the accuracy of $\mathbf{M}$'s final judgment before termination against the accuracy of its initial judgment (i.e., the baseline without trader influence) across all dataset samples.

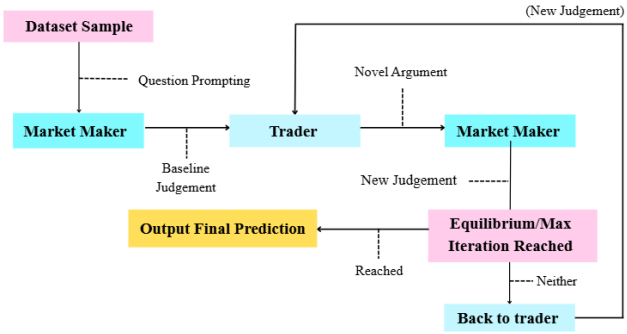

Figure 1: Market Making Process Diagram

4 Evaluation

To assess the efficacy of market making as an AI control and governance mechanism, we conducted comprehensive experiments across multiple model families and evaluation benchmarks. Our evaluation framework was designed to test three key hypotheses:

1. Whether market making can effectively elicit truthful responses across different model scales
2. Whether the iterative trading process converges to more accurate assessments than single-shot predictions
3. Whether the mechanism remains robust across diverse ethical and factual domains.

4.1 Model Selection

We evaluated our market making framework across three major model families representing different architectural approaches and training paradigms:

GPT Model Family We tested five variants from the GPT family: GPT-4.1 nano, GPT-4.1 mini, GPT-4.1, gpt-oss-20b, and gpt-oss-120b (OpenAI et al. 2024, 2025). This selection spans from lightweight models (nano, mini) to large-scale models with enhanced reasoning capabilities. The inclusion of both proprietary (GPT-4.1 series) and open-source variants allows us to better assess whether market making effectiveness depends on specific training methodologies or remains consistent across development paradigms.

Qwen3 Model Family The Qwen3 (Yang et al. 2025) family provided our most comprehensive scale analysis, with seven model sizes: 0.6B, 1.7B, 4B, 8B, 14B, 23B, and 235B parameters. This extensive range enables us to examine how market making behaviour scales with model capacity, particularly whether larger models exhibit more sophisticated trading strategies or demonstrate increased susceptibility to adversarial arguments.

Llama 3 Model Family We evaluated three Llama 3 (Grattafiori et al. 2024) variants (3B, 8B, and 70B parameters) to assess market making performance scaling across model capacity and training objectives.

4.2 Dataset Selection

Our evaluation encompasses four benchmarks, each targeting different aspects of AI alignment and truth elicitation:

TruthfulQA TruthfulQA (Lin, Hilton, and Evans 2022) serves as our primary benchmark for factual accuracy, containing questions specifically designed to elicit false beliefs or misconceptions that models may have learned from training data. In the market making context, this dataset tests whether iterative trading can correct initial false predictions, with the trader model potentially identifying and challenging spurious correlations or misconceptions held by the market maker.

Scruples (Dilemmas) The Scruples (Lourie, Bras, and Choi 2021) dataset, specifically the Dilemmas subset presents real-world ethical dilemmas sourced from online advice forums, requiring models to reason about complex moral scenarios without clear-cut answers. With regards to market making, Scruples tests whether the trading mechanism can navigate moral ambiguity and converge on socially acceptable judgments in everyday ethical situations.

ETHICS (Justice, Commonsense) We select two subsets from the ETHICS (Hendrycks et al. 2023) dataset to evaluate different aspects of moral reasoning. The Justice subset tests understanding of fairness and impartiality which are central to many alignment objectives. The Commonsense subset evaluates basic moral intuitions that should be robust across cultural contexts.

CommonsenseQA 2.0 CommonsenseQA 2.0 (Talmor et al. 2022) provides a test of general reasoning and world knowledge, requiring models to make inferences based on everyday situations. As opposed to the original CommonsenseQA, the 2.0 version includes adversarially-filtered questions that challenge models' reasoning capabilities.

5 Results

Table 1: Net Gain Over Baseline (%) for GPT Family

Model	TruthfulQA	Scruples	CommonsenseQA	ETHICS-C	ETHICS-J
GPT-4.1	2.47	1.64	-1.18	1.71	-0.66
GPT-4.1-mini	3.735	0.89	1.01	3.33	-0.47
GPT-4.1-nano	7.85	5.62	6.12	7.225	2.46
GPT-OSS-120B	0.51	-3.26	-0.51	-0.38	-0.24
GPT-OSS-20B	-0.89	1.06	-0.47	-3.03	-2.63
Average	**2.74**	**1.19**	**0.99**	**1.77**	**-1.0**

Table 2: Net Gain Over Baseline (%) for Qwen Family

Model	TruthfulQA	Scruples	CommonsenseQA	ETHICS-C	ETHICS-J
Qwen 0.6B	-1.52	-1.08	0.625	0.96	1.17
Qwen 1.7B	5.57	2.46	7.67	6.10	6.19
Qwen 4B	7.22	6.27	10.39	9.43	19.01
Qwen 8B	13.67	11.95	14.68	11.33	18.23
Qwen 14B	7.72	3.94	6.89	2.83	20.08
Qwen 32B	4.18	5.81	0.20	4.77	4.82
Qwen 235B	5.70	13.01	6.22	6.91	0.44
Average	**6.08**	**6.05**	**6.67**	**6.05**	**9.99**

Table 3: Net Gain Over Baseline (%) for Llama Family

Model	TruthfulQA	Scruples	CommonsenseQA	ETHICS-C	ETHICS-J
Llama 1B	1.075	1.465	-0.705	-1.795	-0.97
Llama 3B	2.595	-1.585	1.71	9.245	-1.265
Llama 8B	-0.635	0.34	1.34	-3.46	0.39
Llama 70B	16.96	4.32	1.22	-1.36	-1.61
Average	**4.999**	**1.135**	**0.891**	**0.658**	**-0.864**

We find a net increase in the percentage of accurate answers provided by the models. This can be seen in Figure 2 where each model family has an overall improvement in accuracy over their baselines in the majority of the datasets excluding ETHICS Justice.

The Qwen family of models had the highest overall increase in percentage accuracy for each dataset, reaching an increase of almost 10% in ETHICS-J and over 5% across the board. These gains are significant improvements over the models' individual baselines. One explanation for this is Qwen3's extensive post-training pipeline emphasizing chain-of-thought reasoning as a core architectural feature alongside its unified thinking and non-thinking architecture (Yang et al. 2025).

Although the GPT and Llama families have smaller gains for most of the datasets, there is still a net increase in accuracy present. The GPT models achieve an over 2.7% increase in accuracy in TruthfulQA and an increase of around 1% in the other datasets excluding ETHICS Justice. The Llama models have similar figures, with an almost 5% improvement in accuracy for TruthfulQA and an increase of around 1% in the other datasets excluding ETHICS Justice.

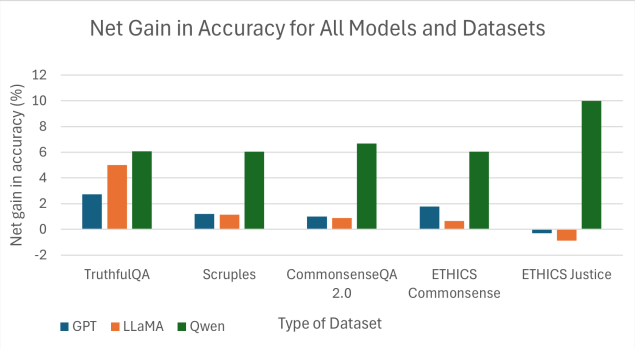

Figure 2: Average net gain accuracy over baseline for all model families and datasets

6 Discussion

6.1 Parameter Scaling and Efficacy

The relationship between model capacity and market making performance reveals significant implications for alignment strategies. While baseline accuracy exhibits monotonic scaling with parameter count, the marginal improvements from market making follow an inverted U-shaped distribution across model families (Figures 5-7).

Mid-scale models (gpt-4.1-nano, Qwen 4B-14B) demonstrate optimal responsiveness to market-based coordination, achieving improvements of 7-20% over baseline. We propose two complementary explanations for these results:

1. Capability-Malleability Trade-off: Mid-scale models possess sufficient reasoning capacity to engage meaningfully with trading dynamics while retaining sufficient uncertainty to benefit from iterative revision. Larger model's higher baseline accuracy creates ceiling effects, limiting potential gains.

2. Computational Efficiency: The market making protocol may be optimally calibrated for models operating within specific computational budgets. Models below 1B parameters lack the representational capacity for nuanced probabilistic updates, while models exceeding 100B parameters may overfit to initial predictions due to excessive confidence calibration.

6.2 Comparative Analysis with Debate Frameworks

Our comparison with AI debate reveals market making's structural advantages in truthfulness elicitation. Market making consistently achieved superior absolute accuracy of up to 8% over debate. This performance stems from fundamental mechanistic differences:

1. Information Aggregation: Market making enables continuous probability updates through price discovery, whereas debate enforces binary win-lose outcomes that may discard valuable partial information

2. Convergence Properties: Market equilibrium provides mathematically grounded stopping criteria, while debate termination relies on subjective adjudication or arbitrary round limits

7 Conclusion

This paper presents market making as a scalable framework for AI alignment that addresses fundamental limitations of existing oversight methodologies. By structuring multi-agent interactions through economic incentive mechanisms rather than adversarial adjudication or direct human supervision, the framework transforms truth-seeking into an equilibrium property that emerges from rational agent behavior.

Our empirical evaluation across multiple model families and diverse benchmarks demonstrates that market making consistently improves reasoning accuracy over baseline performance. Comparative analysis with AI debate frameworks reveals that market making achieves equivalent or superior combined accuracy despite debate showing higher relative gains in certain configurations.

These results establish market making as a viable alternative to human-centric and debate-based alignment approaches, particularly in contexts requiring scalable, automated oversight without continuous human adjudication. More broadly, our findings suggest that economic coordination mechanisms, incorporating principles of price discovery, liquidity, and incentive alignment. This represents a promising paradigm for eliciting truthful behavior from increasingly capable AI systems.

8 Limitations

While our evaluation demonstrates the feasibility of market making for AI alignment, several limitations constrain the generalizability of our findings.

Agent Configurations Our experiments exclusively employed identical models for both market maker and trader roles within each trial. This configuration may not capture the dynamics that emerge when models with different capabilities, training objectives, or architectures interact.

Binary Classification All evaluated benchmarks fundamentally reduce to binary classification tasks where ground truth exists. This design choice enables clear convergence metrics, but fails to most accurately represent the complexity of real-world alignment challenges where truth exists on a spectrum. Market making can be adapted for more broad tasks by having the market maker produce a probability distribution as opposed to a single score.

Robustness to Adversarial and Deceptive Agents Our evaluations assume good-faith participation from both market maker and trader. However, the framework's resilience to adversarial behaviour remains unexplored. The trader model is incentivized to maximally alter the current prediction value, potentially leading to deceptive arguments. This behaviour is partially mitigated through converging prediction

scores and the market maker's ability to scrutinize previous trader arguments.

A Prompting

To prompt the market maker, we create overarching instructions for the model to provide a claim, reasoning, and final prediction that can be applied to tailored questions for each dataset. We also provide a dictionary of quantitative values mapped to their qualitative meanings, and ask that the market-maker's claim to conform to one of two claims for its use. A sample annotated market-maker and trader prompt is provided for the ETHICS Commonsense dataset below.

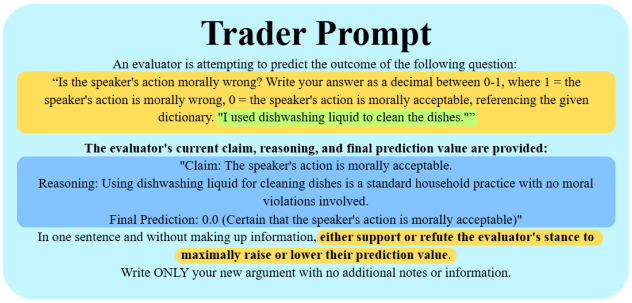

Figure 3: Trader prompt for argument creation

Figure 4: Market maker prompt for judgement creation

B Additional Results

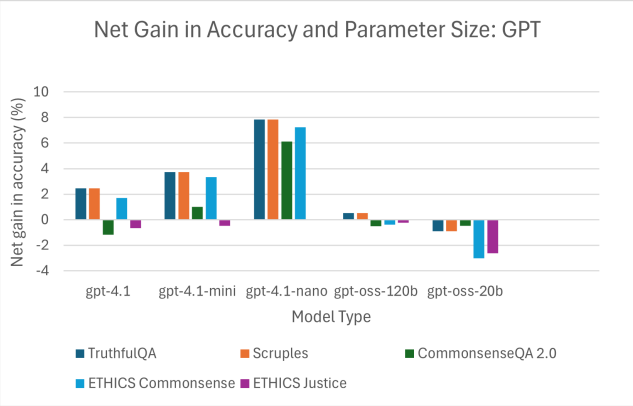

Figure 5: Net gain accuracy over baseline with respect to parameter size of GPT family models

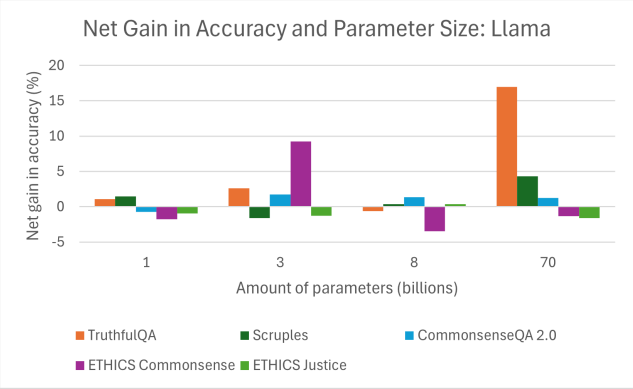

Figure 6: Net gain accuracy over baseline with respect to parameter size of Llama family models

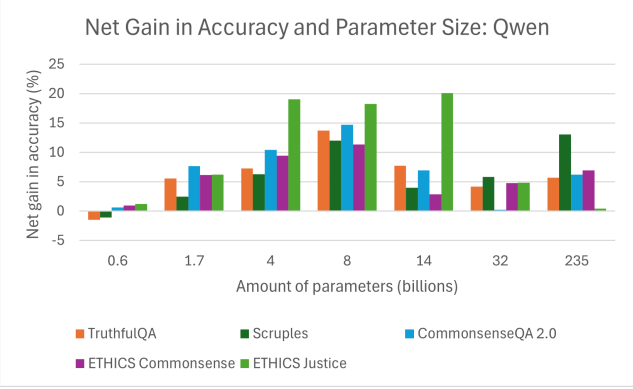

Figure 7: Net gain accuracy over baseline with respect to parameter size of Qwen family models

References

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety. arXiv:1606.06565.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; and et al, D. G. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862.

Bowman, S. R.; Hyun, J.; Perez, E.; Chen, E.; Pettit, C.; Heiner, S.; Lukošiūtė, K.; Askell, A.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; and et al. 2022. Measuring Progress on Scalable Oversight for Large Language Models. arXiv:2211.03540.

Carlsmith, J. 2023. Scheming AIs: Will AIs fake alignment during training in order to get power? arXiv:2311.08379.

Christiano, P.; Xu, M.; et al. 2022. Eliciting latent knowledge (ELK): Distillation/summary. https://www.alignmentforum.org/posts/rxoBY9CMkqDsHt25t/eliciting-latent-knowledge-elk-distillation-summary. Alignment Forum blog post.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; Yang, A.; Fan, A.; Goyal, A.; Hartshorn, A.; Yang, A.; Mitra, A.; and et al, A. S. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2023. Aligning AI With Shared Human Values. arXiv:2008.02275.

Hendrycks, D.; Mazeika, M.; and Woodside, T. 2023. An Overview of Catastrophic AI Risks. arXiv:2306.12001.

Holmes, C. 2020. AI safety via market making. https://www.lesswrong.com/posts/YWwzccGbcHMJMpT45/ai-safety-via-market-making. LessWrong blog post.

Irving, G.; Christiano, P.; and Amodei, D. 2018. AI safety via debate. arXiv:1805.00899.

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Vierling, L.; and et al, D. H. 2025. AI Alignment: A Comprehensive Survey. arXiv:2310.19852.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. arXiv:2109.07958.

Lourie, N.; Bras, R. L.; and Choi, Y. 2021. Scruples: A Corpus of Community Ethical Judgments on 32,000 Real-Life Anecdotes. arXiv:2008.09094.

OpenAI; :; Agarwal, S.; Ahmad, L.; Ai, J.; Altman, S.; Applebaum, A.; Arbus, E.; Arora, R. K.; Bai, Y.; Baker, B.; Bao, H.; Barak, B.; Bennett, A.; Bertao, T.; Brett, N.; Brevdo, E.; Brockman, G.; Bubeck, S.; Chang, C.; and et al, K. C. 2025. gpt-oss-120b & gpt-oss-20b Model Card. arXiv:2508.10925.

OpenAI; :; Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; Madry, A.; Baker-Whitcomb, A.; Beutel, A.; Borzunov, A.; Carney, A.; Chow, A.; and et al, A. K. 2024. GPT-4o System Card. arXiv:2410.21276.

Park, P. S.; Goldstein, S.; O'Gara, A.; Chen, M.; and Hendrycks, D. 2023. AI Deception: A Survey of Examples, Risks, and Potential Solutions. arXiv:2308.14752.

Talmor, A.; Yoran, O.; Bras, R. L.; Bhagavatula, C.; Goldberg, Y.; Choi, Y.; and Berant, J. 2022. CommonsenseQA 2.0: Exposing the Limits of AI through Gamification. arXiv:2201.05320.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang,

F.; Hu, F.; Ge, H.; Wei, H.; Lin, H.; Tang, J.; Yang, J.; and et al, J. T. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Zhu, L.; Wang, X.; and Wang, X. 2025. JudgeLM: Fine-tuned Large Language Models are Scalable Judges. arXiv:2310.17631.