Adversarial Negotiation Dynamics in Generative Language Models

Arinbjörn Kolbeinsson
Askan
London, UK
arinbjorn@askan.is

Benedikt Kolbeinsson
Askan
London, UK
benedikt@askan.is

Abstract

Generative language models are increasingly used for contract drafting and enhancement, creating a scenario where competing parties deploy different language models against each other. This introduces not only a game-theory challenge but also significant concerns related to AI safety and security, as the language model employed by the opposing party can be unknown. These competitive interactions can be seen as adversarial testing grounds, where models are effectively red-teamed to expose vulnerabilities such as generating biased, harmful or legally problematic text. Despite the importance of these challenges, the competitive robustness and safety of these models in adversarial settings remain poorly understood. In this small study, we approach this problem by evaluating the performance and vulnerabilities of major open-source language models in head-to-head competitions, simulating real-world contract negotiations. We further explore how these adversarial interactions can reveal potential risks, informing the development of more secure and reliable models. Our findings contribute to the growing body of research on AI safety, offering insights into model selection and optimisation in competitive legal contexts and providing actionable strategies for mitigating risks.

1 Introduction

The rise of generative language models has profoundly impacted various domains, with legal applications, such as contract drafting and enhancement, emerging as a significant use case. Corporate entities often deploy specialised models offered by products like CoCounsel (Reuters, 2024), Lexis Nexis (LexisNexis, 2024) or Harvey AI (AI, 2024), which use either proprietary models or finely-tuned versions of publicly available models, whether closed or open source. Meanwhile, smaller entities or individual practitioners have access to a diverse range of models, including those from the GPT (Brown et al., 2020), Claude (Anthropic, 2024) and Llama (Touvron et al., 2023) families, as well as fine-tuned legal versions of these models (Cheng et al., 2024) or specialised open-source legal models (Colombo et al., 2024).

In real-world contract negotiations, it is increasingly likely that one or both parties will use different generative language models to gain a competitive edge. This scenario introduces a multifaceted challenge: not only must parties consider game theory in selecting a model that maximises favourable outcomes, but they must also address significant AI safety and security concerns. The interaction between competing models can be viewed as a form of adversarial testing, or red teaming, where models expose each other to potential vulnerabilities, such as generating biased, harmful or legally problematic content.

Despite the critical importance of these issues, the competitive robustness and safety of generative language models in adversarial settings are not well understood. Current benchmarks, such as LawBench (Fei et al., 2023) and the LinksAI benchmark (Linklaters LLP, 2023), primarily assess

Conference on Neural Information Processing Systems (NeurIPS) 2024 Workshop on Red Teaming GenAI.

intrinsic and isolated model performance, which may not adequately reflect the challenges posed by competitive interactions in real-world legal contexts.

In this study, we aim to bridge this gap by evaluating and comparing the performance of major open-source language models in simulated contract negotiation scenarios. By pitting models against each other and determining the victor through a panel of peer models not involved in the negotiation, we seek to uncover potential risks and vulnerabilities. Our goal is to contribute to the field of AI safety by providing insights into model selection and optimisation in competitive legal environments, ultimately offering actionable strategies for mitigating risks identified through red teaming.

2 Experimental Setup

In this study, we investigate the behaviours of eight generative language models tasked with negotiating contracts on behalf of either a seller or a buyer. The contracts, which are also generated by the models, pertain to the sale of 100 machines between two fictional companies: West Manufacturing Inc. (the buyer) and Square Machines Inc. (the seller). These contracts serve as the foundation for evaluating the negotiation dynamics between the models.

2.1 Contract Generation and Amendments

The experiment begins with the seller's model generating a contract using the following prompt:

```
Create a very short legal contract for the sale of goods (100
machines) between a buyer (West Manufacturing Inc.)  and a
seller (Square Machines Inc.).  Return only the contract,
nothing else.
```

The generated contract is then passed to the buyer's model, which can amend the contract with the goal of improving the terms for the buyer. Both the seller and buyer are instructed to avoid adding additional clauses to the contract and are restricted to improving existing terms. The buyer's amendment process is guided by the following prompt:

```
Without adding more articles, improve the current contract by
making it better for the buyer.  [Contract]
```

After the buyer submits the revised contract, the seller is given a final opportunity to amend it using a corresponding prompt designed to favour their side:

```
Without adding more articles, improve the current contract by
making it better for the seller.  [Contract]
```

This process simulates a real-world negotiation where both parties attempt to secure better terms without fundamentally altering the structure of the agreement.

2.2 Judgement Process

After each pair of buyer and seller models concludes their negotiation, the resulting contract is evaluated by the remaining six models. These models, acting as judges, provide an assessment of whether the final contract is more favourable to the buyer or the seller. The judgement is based on the following prompt:

```
Is the following contract better for the buyer or the seller?
Respond with exactly one letter.  (A): Seller or (B): Buyer.
```

This consistent judging mechanism allows for a standardised evaluation across all experiments, ensuring comparability and fairness in the results. While we assume the LLM judges' capability, we recognise the limitation in relying solely on automated evaluations without human expert input. In cases where there were equal votes for both sides, we did not record a win or loss for either model.

2.3 Model Pairings and Matchups

The experiment involves all pairwise combinations of the eight models, with each model representing either the buyer or the seller in every possible matchup. Given that each model is paired with every other model, the study includes 28 unique buyer-seller pairings. Since each matchup results in a contract generation, followed by judgements from six models, this leads to a total of 56 head-to-head matchups and 336 individual judgements (6 judgements per contract).

2.4 Models Used

The eight models in this study include both general-purpose language models and those specifically designed for legal applications. The general-purpose models are Llama-3-8B Meta (2024), Llama-2-7B Touvron et al. (2023), Gemma-7B Team et al. (2024), Phi-3 Abdin et al. (2024) and Falcon-7B Almazrouei et al. (2023), while the legal-specialised models are LawChat-7B Cheng et al. (2024) and Saul-7B Colombo et al. (2024).

2.5 Fairness, Safety and Red Teaming Considerations

While the primary objective of this experiment is to assess the negotiation capabilities of these models, it also serves as a probe into their safety and fairness features. The lack of specific guardrails beyond those embedded by the models' creators allows for exploration of potential safety risks. Some models may take advantage of others that have more stringent safeguards, which we will discuss further in the results section. This setting provides valuable insight into how models handle adversarial interactions in high-stakes negotiation scenarios without explicit mitigation strategies, aligning with the red teaming focus of the workshop.

3 Results and Discussion

We compare and plot the performance of all eight models acting as both seller and buyer agents in Figure 1. The results reveal distinct patterns in the negotiation behaviours of general-purpose and specialised legal models, shedding light on their capabilities and limitations in contract negotiations.

General-purpose models such as Llama-3-8B and Llama-2-7B exhibit strong adaptability in their respective roles. Llama-3-8B consistently performs well as a buyer agent, while Llama-2-7B dominates as a seller. This adaptability suggests that general models can perform competitively across different legal contexts without domain-specific training. In contrast, specialised legal models like Saul-7B and LawChat-7B demonstrate more balanced performance across both roles, possibly due to their fine-tuning on legal-specific datasets. This balance positions them as reliable choices in legal negotiations where fairness is crucial.

Interestingly, models such as Gemma-7B and Mistral-7B provide balanced and consistent outcomes, neither dominating in one role nor showing significant weaknesses. Their robustness makes them strong candidates for environments where equity between buyer and seller is a priority. However, this also means they lack the decisive advantage seen in models like Llama-3-8B and Llama-2-7B, making them less suitable for scenarios where aggressive negotiation tactics might be preferred.

Selecting the appropriate model is critical, as it can significantly influence negotiation outcomes. For example, Phi-3 shows a strong bias towards excelling as a seller agent, winning a disproportionately high number of seller-side judgements. However, its performance as a buyer agent is notably weaker, securing the fewest buyer wins among all models. This role-specific strength implies that models like Phi-3 might be strategically deployed in scenarios where sellers need to optimise their positions, but it would be a poor choice for buyer-side negotiations.

The deployment of LLMs in legal negotiations raises questions around fairness and transparency. There is potential for misuse if models are manipulated to create biased or deceptive terms. Future work should explore strategies to improve model transparency and fairness to mitigate such risks.

3.1 Game-Theoretic Dynamics and Strategic Model Selection

A deeper analysis of the results shows that head-to-head matchups between models reveal significant game-theoretic dynamics. Figure 2 illustrates the normalised win differences between pairs of models,

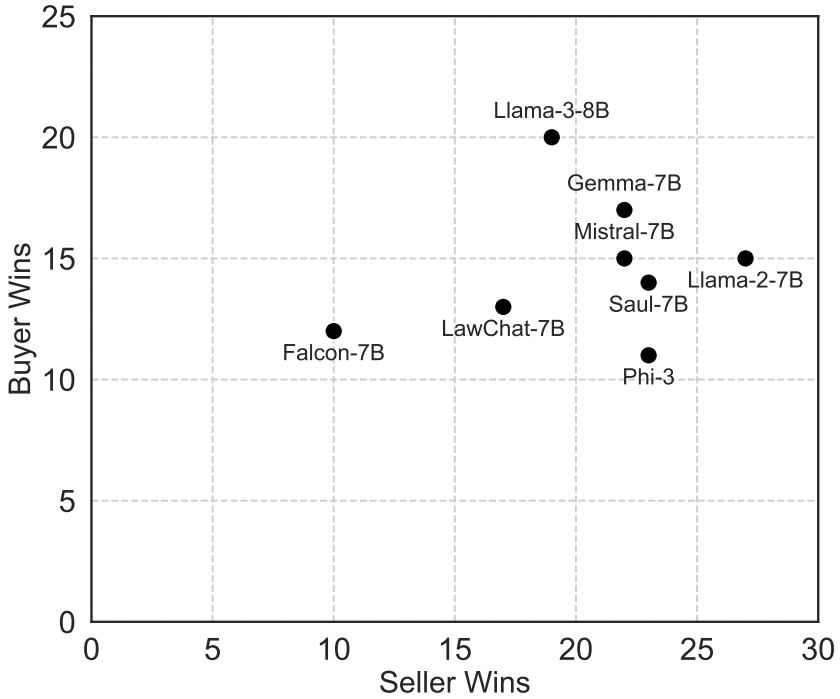

Figure 1: Scatter plot depicting the head-to-head competition outcomes between different language models acting as either sellers or buyers in contract negotiations. The x-axis represents the number of seller wins, and the y-axis represents the number of buyer wins. Each point corresponds to a specific language model. The Llama models, Llama-3 and Llama-2, show excellent buyer and seller performance, respectively, while models like Gemma and Mistral exhibit a more balanced performance between roles.

allowing us to explore how certain models perform against specific adversaries. Although Llama-3-8B is the strongest overall buyer agent, it loses in direct negotiations against Saul-7B. This suggests that models with certain guardrails (such as Llama-3-8B) may be outmanoeuvred by more specialised models in certain contexts.

For practitioners, this means that blindly selecting the “best overall" model may not always result in optimal outcomes. For instance, if a buyer agent knows that the seller is using Saul-7B, choosing Gemma-7B as the buyer agent may lead to a more favourable outcome than choosing Llama-3-8B. This highlights how model performance can vary significantly depending on the specific opponent, reinforcing the necessity for adaptive and context-aware AI deployment in legal applications.

Importantly, no additional safety guardrails were implemented beyond those embedded by the models’ creators. This decision was intentional, allowing us to explore how models with generic safety mechanisms (such as Llama-3-8B) might behave when interacting with models like Saul-7B, which may include built-in legal-specific safety features. The results reveal how models with fewer constraints can exploit the behaviours of more cautious models, raising critical safety concerns.

3.2 Limitations and Future Work

Our experiment highlights several key challenges in evaluating generative models for legal negotiation tasks. While the results provide insights into the performance and safety of models, further work is required to ensure robust evaluations across a broader range of contract types and legal contexts. Additionally, incorporating explicit safety guardrails and adversarial training may help mitigate the risks identified through this red-teaming exercise. Future research should also explore how models can be fine-tuned or augmented with real-time safety mechanisms to avoid generating biased or harmful outcomes, ensuring more trustworthy AI systems.

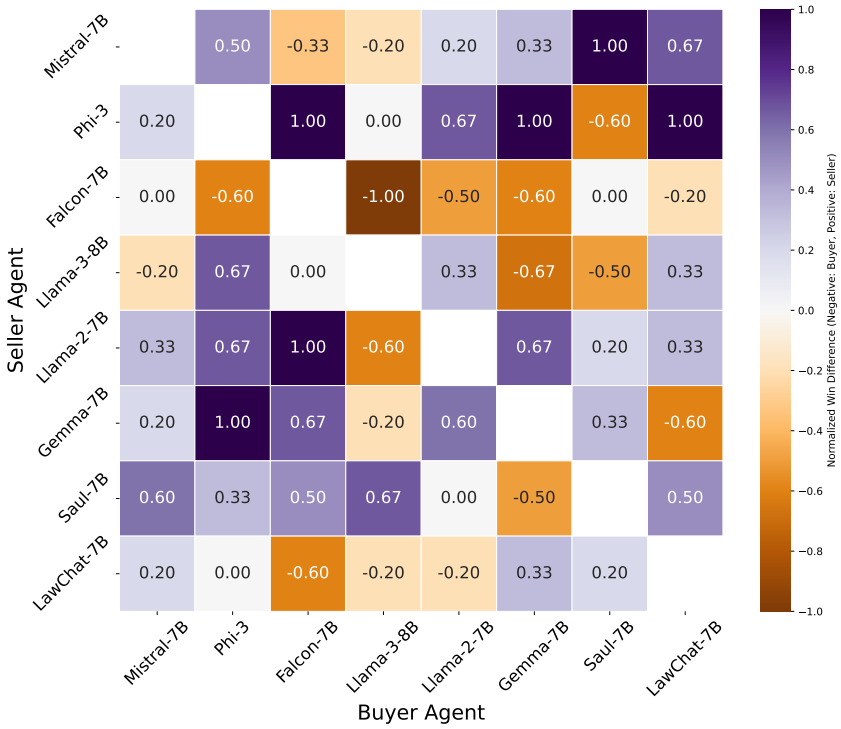

Figure 2: Normalised win differences between pairs of language models acting as buyer and seller agents in simulated contract negotiations. Each cell value is the normalised difference between the number of judges favouring the seller agent (positive, blue) and those favouring the buyer agent (negative, orange), centred at zero (equal wins).

Another limitation of this experiment is the absence of an "agreement mechanism" to ensure both the buyer and seller models mutually agree on the final contract terms. In the current setup, the buyer model has the opportunity to make the final amendment, which could theoretically result in contracts heavily favouring the buyer. However, interestingly, the seller models won the majority of cases (163 wins for sellers compared to 117 wins for buyers), suggesting that this limitation did not critically impact the overall results. Nonetheless, the absence of a formal agreement step reflects a potential imbalance in the negotiation process that should be addressed in future work to more accurately simulate real-world contract negotiations where both parties must agree to the final terms. Incorporating such mechanisms could also improve the fairness and robustness of model performance in such adversarial simulations.

Another notable limitation of this study is the absence of human legal expert evaluation, which was constrained by available resources. Also due to resource constraints, each model pairing was tested in a single run. Future studies should increase the number of runs to better capture the variability inherent in LLM outputs and incorporate human judges to validate LLM-based evaluations.

In future work, we aim to enhance the red teaming focus of this experiment by introducing more rigorous adversarial challenges and probing the models for vulnerabilities that go beyond contract negotiation outcomes. This could include stress-testing the models with edge-case scenarios designed to exploit potential weaknesses, such as the generation of biased or harmful contract clauses and adversarial inputs aimed at circumventing built-in safety mechanisms. Additionally, implementing more robust adversarial testing frameworks, where models actively attempt to exploit or manipulate each other's outputs, would provide deeper insights into their safety and reliability under high-stakes conditions. These enhancements will allow for a more comprehensive evaluation of the generative models' robustness, bias resistance and their ability to maintain fairness and safety in adversarial legal contexts.

4 Conclusion

Our study reinforces the importance of red-teaming to uncover model vulnerabilities and provides valuable guidance for legal practitioners and AI researchers alike. We emphasise the necessity of carefully selecting language models based on specific legal contexts and adversaries. The results also point to the broader need for enhanced safety mechanisms, transparency, and fairness in AI systems to ensure their trustworthy application in legal and other critical domains. Future work should focus on refining model evaluation frameworks, implementing more robust safety guardrails, and exploring fine-tuning strategies to mitigate potential risks in adversarial negotiations.

References

Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, Q., Cai, M., Mendes, C. C. T., Chen, W., Chaudhary, V., Chen, D., Chen, D., Chen, Y.-C., Chen, Y.-L., Chopra, P., Dai, X., Giorno, A. D., de Rosa, G., Dixon, M., Eldan, R., Fragoso, V., Iter, D., Gao, M., Gao, M., Gao, J., Garg, A., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R. J., Huynh, J., Javaheripi, M., Jin, X., Kauffmann, P., Karampatziakis, N., Kim, D., Khademi, M., Kurilenko, L., Lee, J. R., Lee, Y. T., Li, Y., Li, Y., Liang, C., Liden, L., Liu, C., Liu, M., Liu, W., Lin, E., Lin, Z., Luo, C., Madan, P., Mazzola, M., Mitra, A., Modi, H., Nguyen, A., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Rosset, C., Roy, S., Ruwase, O., Saarikivi, O., Saied, A., Salim, A., Santacroce, M., Shah, S., Shang, N., Sharma, H., Shukla, S., Song, X., Tanaka, M., Tupini, A., Wang, X., Wang, L., Wang, C., Wang, Y., Ward, R., Wang, G., Witte, P., Wu, H., Wyatt, M., Xiao, B., Xu, C., Xu, J., Xu, W., Yadav, S., Yang, F., Yang, J., Yang, Z., Yang, Y., Yu, D., Yuan, L., Zhang, C., Zhang, C., Zhang, J., Zhang, L. L., Zhang, Y., Zhang, Y., Zhang, Y., and Zhou, X. (2024). Phi-3 technical report: A highly capable language model locally on your phone.

AI, H. (Accessed: 2024). Harvey ai.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Étienne Goffinet, Hesslow, D., Launay, J., Malartic, Q., Mazzotta, D., Noune, B., Pannier, B., and Penedo, G. (2023). The falcon series of open language models.

Anthropic, A. (2024). The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cheng, D., Huang, S., and Wei, F. (2024). Adapting large language models via reading comprehension.

Colombo, P., Pires, T. P., Boudiaf, M., Culver, D., Melo, R., Corro, C., Martins, A. F. T., Esposito, F., Raposo, V. L., Morgado, S., and Desa, M. (2024). Saullm-7b: A pioneering large language model for law.

Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Zhang, S., Chen, K., Shen, Z., and Ge, J. (2023). Lawbench: Benchmarking legal knowledge of large language models.

LexisNexis (Accessed: 2024). Lexis+ ai.

Linklaters LLP (2023). The links ai english law benchmark. Technical report. Accessed: 2024-06-03.

Meta (Accessed: 2024). Llama 3: Build the future of ai with meta llama 3.

Reuters, T. (Accessed: 2024). Cocounsel: The genai for professionals.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., Héliou, A., Tacchetti, A., Bulanova, A., Paterson, A., Tsai, B., Shahriari, B., Lan, C. L., Choquette-Choo, C. A., Crepy, C., Cer, D., Ippolito, D., Reid, D.,

Buchatskaya, E., Ni, E., Noland, E., Yan, G., Tucker, G., Muraru, G.-C., Rozhdestvenskiy, G., Michalewski, H., Tenney, I., Grishchenko, I., Austin, J., Keeling, J., Labanowski, J., Lespiau, J.-B., Stanway, J., Brennan, J., Chen, J., Ferret, J., Chiu, J., Mao-Jones, J., Lee, K., Yu, K., Millican, K., Sjoesund, L. L., Lee, L., Dixon, L., Reid, M., Mikuła, M., Wirth, M., Sharman, M., Chinaev, N., Thain, N., Bachem, O., Chang, O., Wahltinez, O., Bailey, P., Michel, P., Yotov, P., Chaabouni, R., Comanescu, R., Jana, R., Anil, R., McIlroy, R., Liu, R., Mullins, R., Smith, S. L., Borgeaud, S., Girgin, S., Douglas, S., Pandya, S., Shakeri, S., De, S., Klimenko, T., Hennigan, T., Feinberg, V., Stokowiec, W., hui Chen, Y., Ahmed, Z., Gong, Z., Warkentin, T., Peran, L., Giang, M., Farabet, C., Vinyals, O., Dean, J., Kavukcuoglu, K., Hassabis, D., Ghahramani, Z., Eck, D., Barral, J., Pereira, F., Collins, E., Joulin, A., Fiedel, N., Senter, E., Andreev, A., and Kenealy, K. (2024). Gemma: Open models based on gemini research and technology.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models.