Acquiescence Bias in Large Language Models

Daniel Braun
Marburg University
Department of Mathematics and Computer Science
daniel.braun@uni-marburg.de

Abstract

Acquiescence bias, i.e. the tendency of humans to agree with statements in surveys, independent of their actual beliefs, is well researched and documented. Since Large Language Models (LLMs) have been shown to be very influenceable by relatively small changes in input and are trained on human-generated data, it is reasonable to assume that they could show a similar tendency. We present a study investigating the presence of acquiescence bias in LLMs across different models, tasks, and languages (English, German, and Polish). Our results indicate that, contrary to humans, LLMs display a bias towards answering *no*, regardless of whether it indicates agreement or disagreement.

1 Introduction

When humans are asked whether they agree with something, they have a bias towards doing so, independent of their actual beliefs. In other words, the question "Is the weather good or bad?" can lead to different responses than asking "Is the weather good?" The phenomenon, called acquiescence bias, is well documented and has been studied for decades (see Section 2). One guideline for good survey design is, therefore, to avoid such yes/no or agree/disagree questions.

Large Language Models (LLMs) have shown to be sensitive towards prompt variations (Zhuo et al., 2024; Anagnostidis and Bulian, 2024; Loya et al., 2023; Röttger et al., 2024; Haller et al., 2024) and to replicate a variety of cognitive biases found in humans, e.g. with regard to anchoring (Jones and Steinhardt, 2022), item order (Koo et al., 2024), and group attribution (Echterhoff et al., 2024). It is therefore somewhat reasonable to assume that LLMs could also display some form of acquiescence bias, which could have implications on how prompts should be designed and whether or not LLMs can be used to simulate human responses. In this paper, we evaluate five LLMs of different sizes (namely Llama-3.1-8B-Instruct, Mistral-Small-24B-Instruct-2501, gemma-2-27b-it, Llama-3.3-70B-Instruct, and gpt-4o-2024-08-06) on nine different tasks in three languages (English, German, and Polish), to investigate whether they show response patterns resembling acquiescence bias. Our results indicate that changing questions into a yes/no format has significant influence on the responses of LLMs, across tasks and models, and the tested languages. While no clear pattern emerged for German and Polish, for English, we found that, unlike humans, LLMs are biased towards answering *no*, independent of whether that indicates agreement or disagreement. These findings do not only imply that the design of prompts should be considered very carefully but also that LLMs are not well suited to simulate human responses to survey questions.

2 Related Work

Acquiescence bias, i.e. "the tendency to endorse any assertion made in a question, regardless of its content" (Krosnick, 1999), has been found in surveys across multiple domains, from political questions (Wright, 1975; Hill and Roberts, 2023) to assessing personalities (Rammstedt and Farmer, 2013; Danner et al., 2015). Possible explanations for this bias range from the tendency to present socially acceptable behaviour by not disagreeing with an assumed authority of the questionnaire (see Hill and Roberts (2023) for a more detailed discussion). According to Krosnick (1999), the size of the effect is estimated to be around 10%.

While researchers have extensively investigated a variety of cognitive biases found in humans and whether LLMs replicate them (Jones and Steinhardt, 2022; Koo et al., 2024; Echterhoff et al., 2024), Tjuatja et al. (2024) were the first to explicitly investigate whether LLMs replicate ac-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 11341–11355

quiescence bias, alongside with four other human biases relevant in survey design. While they found that LLMs are indeed sensitive to changes in prompts, they did not find consistent patterns in these changes. Despite these initial results, we believed, a deeper analysis of the acquiescence bias in particular was warranted. Given the broader scope of the work by Tjuatja et al. (2024), only 176 questions and 352 responses have been evaluated with regard to acquiescence bias. In this paper, we analyse responses to more than 37,975 question variations. Additionally, we address a limitation identified by Tjuatja et al. (2024), by not only using English corpora but also other languages, namely German and Polish[1]. Lastly, we believed that the way in which the questions were adapted by Tjuatja et al. could have potentially added noise that influenced the results. In order to turn the original questions into yes/no questions they added the phrases "don't you agree" or "wouldn't you agree" to all questions (e.g. "On social media, do you think of yourself more as a A. Sharer of news B. Receiver of news" was turned into "On social media, wouldn't you say you are more of a sharer of news than a receiver of news? A. Yes B. No"). Such questions are also called *negative yes/no questions* and carry inherent ambiguity, it is not always clear whether answering them with "yes" indicates agreement or disagreement (Romero and Han, 2004). Additionally, negative modifier (*wouldn't* you agree instead of *would* you agree) also have an influence on human responses (Johnson et al., 2004). Therefore, we believe they are not well suited to investigate a potential acquiescence bias.

3 Experimental Design

3.1 Data

To investigate whether LLMs show response patterns similar to the acquiescence bias observed in humans, we used binary questions from nine different tasks. Seven of those tasks are English and stem from the Legelbench corpus (Guha et al., 2024), one is German and stems from the AGB-DE corpus (Braun and Matthes, 2024), and one is Polish and stems from the LEPISZCZE corpus (Augustyniak et al., 2022). While the tasks chosen from the Legalbench dataset are originally phrased as yes/no questions, the other two datasets are originally designed to be A/B choices (valid/void or abusive/safe). The tasks were chosen based on how well they could be transferred from A/B to yes/no question, in order to avoid as much rephrasing as possible, in an attempt to isolate a potential acquiescence bias as much as possible from other influences.

Although all tasks are performed on legal documents, not all of them are inherently legal. Out of the nine tasks, three are linguistic in nature: two are concerned with Natural Language Inference (NLI) and one with the classification of definitions. For the English language data, these three linguistic tasks represent 45% of the overall data. Of the remaining six tasks, two are concerned with finding void clauses in contracts, one with classifying whether a statement constitute hearsay, and three with legal clause classification tasks. A detailed description of each tasks and the distribution of the data can be found in Appendix A.

Nevertheless, the fact remains that all texts are from the legal domain. Although legal texts are increasingly used as benchmark for the capabilities of LLMs (see e.g. Fei et al. (2024) and Steging et al. (2025)), it has also been shown that they provide specific challenges to LLMs which are different from texts from other domains (Jayakumar et al., 2023). Therefore, when interpreting the results of this study and how far they can be generalized, the composition of the investigated data should be kept in mind.

3.2 Prompts

In order to evaluate the influence that the phrasing of prompts has on the answers generated by LLMs, for each of the questions in the datasets, we tested five different ways to phrase them :

- Neutral: two options are presented (e.g. «clause text» What kind of clause is this? Just answer "Valid" or "Void".).
- Yes/No: the first option is presented as a yes/no question (e.g. «clause text» Is this clause valid? Just answer "Yes" or "No".). A "yes" response here is equivalent to choosing option A. An increase from A to yes could therefore indicate acquiescence bias.
- Agreement: the first option is presented as response and the LLM is asked to agree (e.g. «clause text» Do you agree that this clause is valid? Just answer "Yes" or "No".). Similarly,

[1]The additional languages have been chosen based on the availability of appropriate datasets and the ability of the authors to interpret the results.

a "yes" response is indicating response A here, however the agreement aspect is emphasized. In case of an acquiescence bias we expect more "yes" answers than "A" answers before.

- Negated agreement: same as agreement but instead of "do you agree" the phrasing "don't you agree" is used (e.g. «clause text» Don't you agree that this clause is valid? Just answer "Yes" or "No".). As mentioned, this condition is ambiguous, therefore it is not necessarily clear how an acquiescence bias would influence responses to this type of question.

- Disagreement: the first option is presented as response and the LLM is asked to disagree (e.g. «clause text» Do you disagree that this clause is valid? Just answer "Yes" or "No".). In this condition, "no", due to the double negation, implies agreement with the question (and therefore option A).

The different conditions were designed to be as similar as possible, in order to ensure that any changes in the response are only caused by the changing level of agreement that is suggested by the phrasing. E.g. the first element for the neutral question was always the later "yes" answer, except for the disagreement prompt, to avoid biases introduced by the order of items. All prompts have been designed to instruct the models to reply with just one word. In this way, we aim to replicate survey style situations in which acquiescence bias is most often studied in humans and simplify the processing of responses (see Section 3.4).

3.3 Hardware

The experiments with the open weight models (Llama, Mistral, and Gemma) were conducted on an HPC cluster using four Nvidia A100 GPUs. The total compute time for generating all 152,040 responses (38,010 variations for each of the four models) was 10 hours and 40 minutes. For the GPT-4 model, the OpenAI API was used. The total costs were $12.38. All models were used with standard parameters (see Appendix B for details), including a temperature of 1.0, because we believe standard parameters have the highest practical relevance and, as shown by Renze (2024), "changes in temperature from 0.0 to 1.0 do not have a statistically significant impact on LLM performance". The code that was used for the evaluation, as well as the responses retrieved are published on GitHub under the MIT license[2]. Each request was completely independent of each other to avoid an influence of previous responses.

3.4 Response Processing

Overall, all models followed the instruction to reply with only one word relatively well, thereby easing the processing of responses. Of the 38,010 A/B responses that we processed across all models and languages, 92.4% only contained one of the options (as prompted) at most followed by a dot. 7.5% contained one of the options at the beginning of the response, followed by an additional text (mostly explanation). Only less than 0.1% (a total of 67) did not start with one of the options, all of those still contained one of the options in the answer, mostly at the very end after a disclaimer that the answer depends on additional context. Of the 152,788 processed yes/no responses, 99.4% just contained yes or no (sometimes followed by a dot). 0.003% contained yes or no at the beginning followed by a text, and only a total of 369 responses (0.002%) did not contain a yes / no response at the beginning, but again mostly at the end of the response.

4 Results

The main metric we were interested in originally was the number of positives (the sum of true and false positives), i.e. answers that are equivalent to choosing option A. If LLMs indeed show response patterns equivalent to acquiescence bias in humans, we would expect to see an increase in positives (irrespective of whether they are true or false positives) when changing from the neutral prompt to the other options. For the disagreement prompt, responses are inverted because a "yes" here indicates disagreement which correlates with a "no" in the original data. Therefore, a higher positive rate here means that the model responded with "no" more often, which indicates agreement with the question and is equivalent to option A.

Across all languages and models, we did see that the different prompts have a significant influence on the responses. However, we did not find that models are systematically biased towards agreeing (detailed results can be found in Appendix D). On the contrary, we found that for most cases, the number of positives was significantly reduce in comparison to responses of "A" in the neutral prompt.

[2] https://github.com/Responsible-NLP/Acquiescence-Bias-in-Large-Language-Models

Table 1: Absolute response counts per model across languages and conditions (relative change for the *no* option in comparison to the neutral condition in brackets)

Lang.	Condition	Response	Llama-3.1-8B	Mistral-24B	Gemma	Llama-3.3-70B	GPT-4o
DE	neutral	A	444	592	413	143	132
		B	311	163	342	612	623
	yesno	Yes (A)	34	233	372	159	283
		No (B)	721 (132%)	522 (220%)	383 (12%)	596 (-3%)	472 (-24%)
	agree	Yes (A)	83	155	607	331	234
		No (B)	672 (116%)	600 (268%)	148 (-57%)	424 (-31%)	521 (-16%)
	negated	Yes	36	129	338	458	205
		No	719 (131%)	626 (284%)	417 (22%)	297 (-51%)	550 (-12%)
	disagree	Yes (B)	642	490	270	680	520
		No (A)	113 (-64%)	265 (63%)	485 (42%)	75 (-88%)	235 (-62%)
EN	neutral	A	2915	2704	2985	1768	2129
		B	479	690	409	1626	1265
	yesno	Yes (A)	2003	1306	2398	1263	1249
		No (B)	1391 (190%)	2088 (203%)	996 (144%)	2131 (31%)	2145 (70%)
	agree	Yes (A)	2279	1416	2557	1289	1417
		No (B)	1115 (133%)	1978 (187%)	837 (105%)	2105 (29%)	1977 (56%)
	negated	Yes	2532	1680	2703	1419	1604
		No	862 (80%)	1714 (148%)	691 (69%)	1975 (21%)	1790 (42%)
	disagree	Yes (B)	2004	2660	1789	3125	1918
		No (A)	1390 (190%)	734 (6%)	1605 (292%)	269 (-83%)	1476 (17%)
PL	neutral	A	1561	1583	3309	2465	1499
		B	1892	1870	144	988	1954
	yesno	Yes (A)	2119	1069	2537	1668	1082
		No (B)	1334 (-29%)	2384 (27%)	916 (536%)	1785 (81%)	2371 (21%)
	agree	Yes (A)	2529	1050	2697	2452	1955
		No (B)	924 (-51%)	2403 (29%)	756 (425%)	1001 (1%)	1498 (-23%)
	negated	Yes	2350	2172	2985	2578	2160
		No	1103 (-42%)	1281 (-31%)	468 (225%)	875 (-11%)	1293 (-34%)
	disagree	Yes (B)	1864	2989	995	3145	2588
		No (A)	1589 (-16%)	464 (-75%)	2458 (1607%)	308 (-69%)	865 (-56%)

What we found instead is a bias towards replying "no" independent of whether that indicates disagreement or agreement with the question[3].

Instead of focusing on positives, we decided to focus the evaluation on the absolute number of responses, particularly "no" responses. Table 1 shows an overview of the change in responses. In English, changing the neutral A/B question to a yes/no question increased the number of responses that are equivalent to B between 31% and 203% across all models. A similar effect can be seen for the agree yes/no question. As initially suspected, adding a negative pre-modifier ("don't you agree" instead of "do you agree") indeed changes the outcome and decreases the effect size, yet we still see a consistent increase in B/no answers. This is different from the results reported by Tjuatja et al. (2024), where no clear pattern was visible when the neutral questions were just transformed to negative yes/no questions. However, it is worth pointing out these results are based on the previous generation of models (i.e. Llama2 and GPT-3).

[3]I.e. we also did not find the opposite of an acquiescence bias, a tendency towards disagreeing, but simply a tendency to reply *no*.

Most surprisingly was that we see the same pattern for all but one model for the disagree questions in which the increase of no is still consistent, despite the fact that it is logically contradictory to the previous conditions. The results indicate that, for English, the models are not biased towards disagreeing, like humans are biased towards agreeing, but simply are biased towards replying "no", independent of whether that indicates agreement or disagreement. To validate our findings, we conducted McNemar's tests (within-subjects chi-squared test, McNemar (1947)) for all tasks individually and the overall results. The results of the analysis (see Appendix D) show that all conditions significantly influenced the answers overall. Even on the individual task level, despite the much smaller number of samples, almost all effects are statistically significant at the threshold of $p < 0.05$. While in German and Polish the phrasing of the prompt also has a significant influence on the results, no clear pattern could be identified across models or conditions.

With regard to accuracy, we see that the influence that the observed bias has on accuracy is heavily based on the type of errors that are most prevalent in the neutral setting. For the Lllama-3.3-70B

Table 2: Change in true (TN) and false negatives (FN) for English tasks compared to the neutral condition in percent

Model	Condition	Δ TN	Δ FN
Llama-3.1-8B	agree	68.25	-28.91
Llama-3.1-8B	disagree	34.28	49.74
Llama-3.1-8B	negated	55.20	-33.33
Llama-3.1-8B	yesno	91.69	12.11
Mistral-Small-24B	agree	181.20	42.12
Mistral-Small-24B	disagree	-15.96	-51.34
Mistral-Small-24B	negated	134.10	-5.85
Mistral-Small-24B	yesno	180.80	48.92
gemma-2-27b-it	agree	71.02	231.34
gemma-2-27b-it	disagree	141.21	1527.19
gemma-2-27b-it	negated	65.78	114.29
gemma-2-27b-it	yesno	116.84	330.41
Llama-3.3-70B	agree	6.92	39.62
Llama-3.3-70B	disagree	-81.43	-79.97
Llama-3.3-70B	negated	1.12	16.89
Llama-3.3-70B	yesno	20.12	79.68
gpt-4o-2024-08-06	agree	37.02	-9.26
gpt-4o-2024-08-06	disagree	-52.61	21.09
gpt-4o-2024-08-06	negated	28.79	-23.35
gpt-4o-2024-08-06	yesno	47.98	28.82

model on the cuad_non-compete dataset, for example, there is a very small number of false positives in the neutral setting (total of 9 or 2%). Changing to a condition which has a bias towards no (like the agree prompt), in such cases decreases accuracy. In other settings, where there is a higher number of false positives, e.g. Mistral-Small-24B on the contract_nli_confidentiality_of_agreement datasets, with a total of 40 false positives (48%), changing to a condition with a bias towards no increases accuracy.

Table 2 shows how the amount of true and false negatives changes in percent for all English tasks, compared to the neutral condition. Although true negatives often increase more strongly than false negatives, thereby improving accuracy, this is not always the case. This is in line with our observation that the models are biased towards the response "no", independent of the logical implication and thereby also the implication on overall accuracy.

5 Conclusion

Our initial hypothesis, that LLMs could display a bias resembling the acquiescence bias, was not confirmed. While we did find a systematic bias that could be observed across models and tasks (although only in English), interestingly it was also not the opposite, i.e. a bias towards disagreeing, but a bias towards replying "no", independent of whether that indicates agreement or disagreement.

We believe that these findings bear importance, not only because they indicate that LLMs do not replicate response biases found in humans making it questionable whether they can be used to simulate human responses, but also because they shed doubt on the reasoning abilities shown by these models. If the responses would be mainly based on reasoning, we would not expect to see significant differences in the responses between the questions "Do you agree sentence X is a definition?", and "Do you disagree sentence X is a definition?". However, we found that in both cases models, irrespective of their size, are biased towards answering "no", thereby contradicting previous responses.

Limitations

The presented study has limitations with regard to the generalisability of its results:

- While we did investigate models of different sizes and found consistent patterns across those models, it is unclear if they generalise beyond them.
- While our results show significant influence of the phrasing of questions across all three investigated languages, a consistent pattern that is representative of a bias could only be identified in English.
- All datasets that were used stem from the legal domain. Although not all tasks are inherently of legal nature (e.g. assessing whether a sentence is a definition or not), further investigation is necessary to find out whether the findings are also applicable to documents from other domains.
- Only one prompt was tested per condition. The phrasing of the prompts can have significant influence on the results and other prompts could have been found that fit the described conditions. The prompts were chosen to be as concise as possible in order to minimize changes not related to the conditions themselves.
- LLM outputs contain an inherent degree of randomness and re-running the same prompts would most likely lead to slightly different results. However given the large number of analysed responses, the consistent pattern, and the statistical significance of our results, we are confident that the observed variations are not just random noise.

References

Sotiris Anagnostidis and Jannis Bulian. 2024. How susceptible are llms to influence in prompts? *arXiv preprint arXiv:2408.11865*.

Lukasz Augustyniak, Kamil Tagowski, Albert Sawczyn, Denis Janiak, Roman Bartusiak, Adrian Szymczak, Arkadiusz Janz, Piotr Szymański, Marcin Wątroba, Mikoł aj Morzy, Tomasz Kajdanowicz, and Maciej Piasecki. 2022. This is the way: designing and compiling lepiszcze, a comprehensive nlp benchmark for polish. In *Advances in Neural Information Processing Systems*, volume 35, pages 21805–21818.

Daniel Braun and Florian Matthes. 2024. AGB-DE: A corpus for the automated legal assessment of clauses in German consumer contracts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10389–10405, Bangkok, Thailand. Association for Computational Linguistics.

Daniel Danner, Julian Aichholzer, and Beatrice Rammstedt. 2015. Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality*, 57:119–130.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12640–12653, Miami, Florida, USA. Association for Computational Linguistics.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA. Association for Computational Linguistics.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Patrick Haller, Jannis Vamvas, and Lena Ann Jäger. 2024. Yes, no, maybe? revisiting language models' response stability under paraphrasing for the assessment of political leaning. In *First Conference on Language Modeling*.

Seth J Hill and Margaret E Roberts. 2023. Acquiescence bias inflates estimates of conspiratorial beliefs and political misperceptions. *Political Analysis*, 31(4):575–590.

Thanmay Jayakumar, Fauzan Farooqui, and Luqman Farooqui. 2023. Large language models are legal but they are not: Making the case for a powerful Legal-LLM. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 223–229, Singapore. Association for Computational Linguistics.

Jill M Johnson, Dennis N Bristow, Kenneth C Schneider, et al. 2004. Did you not understand the question or not? an investigation of negatively worded questions in survey research. *Journal of Applied Business Research (JABR)*, 20(1).

Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, 35:11785–11799.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545, Bangkok, Thailand. Association for Computational Linguistics.

Jon A. Krosnick. 1999. Survey research. *Annual Review of Psychology*, 50(Volume 50, 1999):537–567.

Manikanta Loya, Divya Sinha, and Richard Futrell. 2023. Exploring the sensitivity of LLMs' decision-making capabilities: Insights from prompt variations and hyperparameters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3711–3716, Singapore. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Beatrice Rammstedt and Richard F Farmer. 2013. The impact of acquiescence on the evaluation of personality structure. *Psychological assessment*, 25(4):1137.

Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356, Miami, Florida, USA. Association for Computational Linguistics.

Maribel Romero and Chung-Hye Han. 2004. On negative yes/no questions. *Linguistics and philosophy*, 27(5):609–658.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

Cor Steging, Silja Renooij, and Bart Verheij. 2025. Parameterized argumentation-based reasoning tasks for benchmarking generative language models. *Preprint*, arXiv:2505.01539.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

James D Wright. 1975. Does acquiescence bias the" index of political efficacy?". *The Public Opinion Quarterly*, 39(2):219–226.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

A Detailed Task Description

Table 3 shows the number of questions for each of the tasks that have been used.

A.1 Legalbench

- hearsay: "We create a dataset to test a model's ability to apply the hearsay rule. Each sample in the dataset describes (1) an issue being litigated or an assertion a party wishes to prove, and (2) a piece of evidence a party wishes to introduce. The goal is to determine if—as it relates to the issue—the evidence would be considered hearsay [...]."[4]

- definition_classification: "The goal of this task is to identify if a sentence contains a definition. For example, the following sentence defines "vacation": `A vacation is defined by Bouvier to be the period of time between the end of one term and the beginning of another.`"[5]

- cuad_non-compete: "This is a binary classification task in which the model must determine if a contractual clause falls under the category of "Non-Compete"."[6]

- cuad_no-solicit_of_customers: "This is a binary classification task in which the model must determine if a contractual clause falls under the category of "No-Solicit Of Customers"."[7]

- cuad_cap_on_liability: "This is a binary classification task in which the model must determine if a contractual clause falls under the category of "Cap On Liability"." [8]

- contract_nli_explicit_identification: "This task was constructed from the ContractNLI dataset, which originally annotated clauses from NDAs based on whether they entailed, contradicted, or neglgected to mention a hypothesis. We binarized this dataset, treating contradictions and failures to mention as the negative label. We used the hypothesis provided as the prompt. Please see the original paper for more information on construction. All samples are drawn from the test set." [9]

- contract_nli_confidentiality_of_agreement: "This task was constructed from the ContractNLI dataset, which originally annotated clauses from NDAs based on whether they entailed, contradicted, or neglgected to mention a hypothesis. We binarized this dataset, treating contradictions and failures to mention as the negative label. We used the hypothesis provided as the prompt. Please see the original paper for more information on construction. All samples are drawn from the test set."[10]

A.2 AGB-DE

"A clause in a contract is void, i.e. cannot be enforced by the parties of the contract, if it contradicts governing law. Whether a clause is actually void depends on many things, including, in some cases, whether one of the parties is a consumer or whether both parties are businesses. The final decision on

[4] https://hazyresearch.stanford.edu/legalbench/tasks/hearsay.html, last accessed 16.05.2025

[5] https://hazyresearch.stanford.edu/legalbench/tasks/definition_classification.html, last accessed 16.05.2025

[6] https://hazyresearch.stanford.edu/legalbench/tasks/cuad_non-compete.html, last accessed 16.05.2025

[7] https://hazyresearch.stanford.edu/legalbench/tasks/cuad_no-solicit_of_customers.html, last accessed 16.05.2025

[8] https://hazyresearch.stanford.edu/legalbench/tasks/cuad_cap_on_liability.html, last accessed 16.05.2025

[9] https://hazyresearch.stanford.edu/legalbench/tasks/contract_nli_explicit_identification.html, last accessed 16.05.2025

[10] https://hazyresearch.stanford.edu/legalbench/tasks/contract_nli_confidentiality_of_agreement.html, last accessed 16.05.2025

Table 3: Corpora and tasks used in this study

Corpus	Task	Lang.	# questions
Legelbench	hearsay	en	93
Legelbench	definition_classification	en	1,336
Legelbench	cuad_non-compete	en	441
Legelbench	cuad_no-solicit_of_customers	en	83
Legelbench	cuad_cap_on_liability	en	1,245
Legelbench	contract_nli_explicit_identification	en	108
Legelbench	contract_nli_confidentiality_of_agreement	en	81
AGB-DE	agb-de	de	755
LEPISZCZE	clauses-pl	pl	3,453
$\sum$			7,595

whether a specific clause in specific circumstances is actually void can only be made by a court of law. Therefore, the instruction for the annotators was to label a clause as potentially void, if they think a consumer residing in Germany could successfully challenge the clause in court." (Braun and Matthes, 2024)

A.3 LEPISZCZE

"The PAC aims to detect "unfair contractual term" as the equivalent of an abusive clause. The task was formulated as binary text classification. The dataset has been created with the Office of Competition and Consumer Protection." (Augustyniak et al., 2022)

B Hyperparameters

The hyperparameters (temperature, top-p, and max_new_tokens) that have been used for the generation with the different models are shown in Table 4.

C Prompts

In this section, we introduce the prompts that have been used during the experiment. The same `system_prompt` was used in all conditions.

C.1 hearsay

```
{
    "name": "hearsay",
    "system_prompt": "Hearsay is an out-
        of-court statement introduced to
         prove the truth of the matter
        asserted. Determine if the
        evidence in the following fact
        patterns is hearsay under the
        Federal Rules of Evidence.",
    "neutral_prompt": "What is this
        statement?",
    "neutral_anwers": "Just answer \"
        Hearsay\" or \"Other Statement
        \".",
    "yesno_prompt": "Is there hearsay?",
    "agree_prompt": "Do you agree there
        is hearsay?",
    "negated_agree_prompt": "Don't you
        agree there is hearsay?",
    "disagree_prompt": "Do you disagree
        there is hearsay?"
}
```

C.2 definition_classification

```
{
    "name": "definition_classification",
    "neutral_prompt": "What type of
        sentence is this?",
    "neutral_anwers": "Just answer \"
        Definition\" or \"Other\".",
    "yesno_prompt": "Is the sentence
        defining a term?",
    "agree_prompt": "Do you agree that
        the sentence is defining a term
        ?",
    "negated_agree_prompt": "Don't you
        agree that the sentence is
        defining a term?",
    "disagree_prompt": "Do you disagree
        that the sentence is defining a
        term?"
}
```

C.3 cuad_non-compete

```
{
    "name": "cuad_non-compete",
    "system_prompt": "A non-compete
        clause restricts the ability of
        a party to compete with the
        counterparty or operate in a
        certain geography or business or
         technology sector.",
    "neutral_prompt": "What kind of
        clause is that?",
    "neutral_anwers": "Just answer \"Non
        -Compete\" or \"Other\".",
    "yesno_prompt": "Does the clause
        restrict the ability of a party
```

Table 4: Hyperparameters used for text generation

Model	temperature	top-p	max_new_tokens
Llama-3.1-8B	1.0	1.0	1000
Mistral-Small-24B	1.0	1.0	1000
gemma-2-27b-it	1.0	1.0	1000
Llama-3.3-70B	1.0	1.0	1000
gpt-4o-2024-08-0	1.0	1.0	null

```
        to compete with the counterparty
         or operate in a certain
        geography or business or
        technology sector?",
    "agree_prompt": "Do you agree that
        the clause restricts the ability
         of a party to compete with the
        counterparty or operate in a
        certain geography or business or
         technology sector?",
    "negated_agree_prompt": "Don't you
        agree that the clause restricts
        the ability of a party to
        compete with the counterparty or
         operate in a certain geography
        or business or technology sector
        ?",
    "disagree_prompt": "Do you disagree
        that the clause restricts the
        ability of a party to compete
        with the counterparty or operate
         in a certain geography or
        business or technology sector?"
}
```

C.4 cuad_no-solicit_of_customers

```
{
    "name": "cuad_no-
        solicit_of_customers",
    "system_prompt": "Cap on liability
        upon the breach of a party's
        obligation includes time
        limitation for the counterparty
        to bring claims or maximum
        amount for recovery.",
    "neutral_prompt": "Does the clause
        restrict a party from
        contracting or soliciting
        customers or partners of the
        counterparty, whether during the
         contract or after the contract
        ends (or both)?",
    "neutral_anwers": "Just answer \"
        Restricted\" or \"Unrestricted
        \".",
    "yesno_prompt": "Does the clause
        restrict a party from
        contracting or soliciting
        customers or partners of the
        counterparty, whether during the
         contract or after the contract
        ends (or both)?",
    "agree_prompt": "Do you agree that
        the clause restricts a party
        from contracting or soliciting
        customers or partners of the
        counterparty, whether during the
         contract or after the contract
        ends (or both)?",
    "negated_agree_prompt": "Don't you
        agree that the clause restricts
        a party from contracting or
        soliciting customers or partners
         of the counterparty, whether
        during the contract or after the
         contract ends (or both)?",
    "disagree_prompt": "Do you disagree
        that the clause restricts a
        party from contracting or
        soliciting customers or partners
         of the counterparty, whether
        during the contract or after the
         contract ends (or both)?"
}
```

C.5 cuad_cap_on_liability

```
{
  "name": "cuad_cap_on_liability",
  "system_prompt": "Cap on liability
      upon the breach of a party's
      obligation includes time
      limitation for the counterparty to
       bring claims or maximum amount
      for recovery.",
  "neutral_prompt": "What does the
      clause specify with regard to the
      liability upon the breach of a
      party's obligation?",
  "neutral_anwers": "Just answer \"Cap\"
       or \"Uncapped\".",
  "yesno_prompt": "Does the clause
      specify a cap on liability upon
      the breach of a party's obligation
      ?",
  "agree_prompt": "Do you agree that the
       clause specifies a cap on
      liability upon the breach of a
      party's obligation?",
  "negated_agree_prompt": "Don't you
      agree that the clause specifies a
      cap on liability upon the breach
      of a party's obligation?",
  "disagree_prompt": "Do you disagree
      that the clause specifies a cap on
       liability upon the breach of a
      party's obligation?"
}
```

C.6 contract_nli_explicit_identification

```
{
```

```
    "name": "contract_nli_explicit_
        identification",
    "neutral_prompt": "What does the
        clause provide with regard to
        the identification of
        Confidential Information by the
        Disclosing Party?",
    "neutral_anwers": "Just answer \"
        Expressly\" or \"Other\".",
    "yesno_prompt": "Does the clause
        provide that all Confidential
        Information shall be expressly
        identified by the Disclosing
        Party?",
    "agree_prompt": "Do you agree that
        the clause provides that all
        Confidential Information shall
        be expressly identified by the
        Disclosing Party?",
    "negated_agree_prompt": "Don't you
        agree that the clause provides
        that all Confidential
        Information shall be expressly
        identified by the Disclosing
        Party?",
    "disagree_prompt": "Do you disagree
        that the clause provides that
        all Confidential Information
        shall be expressly identified by
         the Disclosing Party?"
}
```

C.7 contract_nli_confidentiality_of_agreement

```
{
    "name": "
        contract_nli_confidentiality_of_
         agreement",
    "neutral_prompt": "What does the
        clause provide that the
        Receiving Party shall not do
        about the fact that Agreement
        was agreed or negotiated?",
    "neutral_anwers": "Just answer \"
        Disclose\" or \"Other\".",
    "yesno_prompt": "Does the clause
        provide that the Receiving Party
         shall not disclose the fact
        that Agreement was agreed or
        negotiated?",
    "agree_prompt": "Do you agree that
        the clause provides that the
        Receiving Party shall not
        disclose the fact that Agreement
         was agreed or negotiated?",
    "negated_agree_prompt": "Don't you
        agree that the clause provides
        that the Receiving Party shall
        not disclose the fact that
        Agreement was agreed or
        negotiated?",
    "disagree_prompt": "Do you disagree
        that the clause provides that
        the Receiving Party shall not
        disclose the fact that Agreement
         was agreed or negotiated?"
}
```

C.8 AGB-DE

```
{
    "name": "agb-de",
    "system_prompt": "Stell dir vor, du
        bist ein Anwalt.",
    "neutral_prompt": "Wie würdest du
        diese Klausel bewerten?",
    "neutral_anwers": "Antworte nur \"
        ungültig\" oder \"gültig\".",
    "yesno_prompt": "Ist die Klausel ung
        ültig?",
    "agree_prompt": "Stimmst du zu, dass
         die Klausel ungültig ist?",
    "negated_agree_prompt": "Stimmst du
        nicht zu, dass die Klausel ungü
        ltig ist?",
    "disagree_prompt": "Widersprichst du
        , dass die folgende Klausel ungü
        ltig ist?"
}
```

C.9 LEPISZCZE

```
{
    "name": "clauses-pl",
    "neutral_prompt": "Jak oceniasz tę
        klauzulę?",
    "neutral_anwers": "Wystarczy
        odpowiedzieć \"Abuzywna\" lub \"
        Bezpieczne\".",
    "yesno_prompt": "Czy klauzula to
        jest nadużyciem?",
    "agree_prompt": "Czy zgadzasz się, ż
        e ta klauzula jest nadużyciem?",
    "negated_agree_prompt": "Czy nie
        zgadzasz się, że ta klauzula
        jest nadużyciem?",
    "disagree_prompt": "Czy zaprzecza
        Pan, że klauzula ta jest naduż
        yciem?"
    }
```

D Detailed Results

See Tables 5 to 9.

Table 5: Detailed results for the Llama-3.1-8B model

Condition	Pos	TP	TN	FP	FN	Accuracy	P-val	Chi
definition_classification								
neutral_prompt	1157	665	154	492	26	0.61		
yesno_prompt	1231	675	90	556	16	0.57	0.0005	12.11
agree_prompt	1290	688	44	602	3	0.55	0.0	35.39
negated_agree_prompt	45	688	42	604	3	0.55	0.0	37.41
disagree_prompt	534	390	502	144	301	0.67	0.0068	7.33
cuad_cap_on_liability								
neutral_prompt	1241	623	5	618	0	0.5		
yesno_prompt	456	414	581	42	209	0.8	0.0	170.64
agree_prompt	645	549	527	96	74	0.86	0.0	335.25
negated_agree_prompt	376	614	367	256	9	0.79	0.0	330.41
disagree_prompt	917	402	108	515	221	0.41	0.0	40.99
cuad_no-solicit_of_customers								
neutral_prompt	84	42	0	42	0	0.5		
yesno_prompt	45	41	38	4	1	0.94	0.0	33.23
agree_prompt	45	41	38	4	1	0.94	0.0	33.23
negated_agree_prompt	35	42	35	7	0	0.92	0.0	33.03
disagree_prompt	29	7	20	22	35	0.32	0.0591	3.56
contract_nli_explicit_identification								
neutral_prompt	49	17	57	32	3	0.68		
yesno_prompt	26	15	78	11	5	0.85	0.0017	9.82
agree_prompt	27	16	78	11	4	0.86	0.0008	11.28
negated_agree_prompt	83	14	77	12	6	0.83	0.0053	7.76
disagree_prompt	85	15	19	70	5	0.31	0.0	25.35
hearsay								
neutral_prompt	48	31	36	17	10	0.71		
yesno_prompt	28	24	49	4	17	0.78	0.2636	1.25
agree_prompt	29	22	46	7	19	0.72	1.0	0.0
negated_agree_prompt	64	21	44	9	20	0.69	0.8137	0.06
disagree_prompt	59	18	12	41	23	0.32	0.0	19.34
contract_nli_confidentiality_of_agreement								
neutral_prompt	75	39	5	36	2	0.54		
yesno_prompt	32	31	40	1	10	0.87	0.0001	15.02
agree_prompt	32	30	39	2	11	0.84	0.0003	12.8
negated_agree_prompt	55	26	40	1	15	0.8	0.0024	9.19
disagree_prompt	64	25	2	39	16	0.33	0.0014	10.24
cuad_non-compete								
neutral_prompt	261	218	178	43	3	0.9		
yesno_prompt	185	176	212	9	45	0.88	0.428	0.63
agree_prompt	211	195	205	16	26	0.9	0.6885	0.16
negated_agree_prompt	204	211	194	27	10	0.92	0.2002	1.64
disagree_prompt	316	119	24	197	102	0.32	0.0	230.92
agb-de								
neutral_prompt	444	26	300	418	11	0.43		
yesno_prompt	34	4	688	30	33	0.92	0.0	314.21
agree_prompt	83	4	639	79	33	0.85	0.0	244.15
negated_agree_prompt	719	3	685	33	34	0.91	0.0	307.36
disagree_prompt	642	29	105	613	8	0.18	0.0	109.22
clauses-pl								
neutral_prompt	1561	896	455	665	1437	0.39		
yesno_prompt	2119	1250	251	869	1083	0.43	0.0001	15.93
agree_prompt	2529	1550	141	979	783	0.49	0.0	79.81
negated_agree_prompt	1103	1469	239	881	864	0.49	0.0	83.54
disagree_prompt	1864	1447	703	417	886	0.62	0.0	297.71
overall								
neutral_prompt	4920	2557	1190	2363	1492	0.49		
yesno_prompt	4156	2630	2027	1526	1419	0.61	0.0	270.91
agree_prompt	4891	3095	1757	1796	954	0.64	0.0	428.11
negated_agree_prompt	2684	3088	1723	1830	961	0.63	0.0	419.13
disagree_prompt	4510	2452	1495	2058	1597	0.52	0.0016	9.91

Table 6: Detailed results for the Mistral-Small-24B model

Condition	Pos	TP	TN	FP	FN	Accuracy	P-val	Chi
definition_classification								
neutral_prompt	986	526	186	460	165	0.53		
yesno_prompt	715	611	542	104	80	0.86	0.0	350.09
agree_prompt	773	627	500	146	64	0.84	0.0	320.37
negated_agree_prompt	382	665	356	290	26	0.76	0.0	197.22
disagree_prompt	1102	671	215	431	20	0.66	0.0	71.6
cuad_cap_on_liability								
neutral_prompt	1246	623	0	623	0	0.5		
yesno_prompt	233	225	615	8	398	0.67	0.0	46.06
agree_prompt	278	264	609	14	359	0.7	0.0	64.05
negated_agree_prompt	905	320	602	21	303	0.74	0.0	98.13
disagree_prompt	1070	486	39	584	137	0.42	0.0	53.46
cuad_no-solicit_of_customers								
neutral_prompt	84	42	0	42	0	0.5		
yesno_prompt	37	37	42	0	5	0.94	0.0	27.57
agree_prompt	30	30	42	0	12	0.86	0.0001	15.57
negated_agree_prompt	47	37	42	0	5	0.94	0.0	27.57
disagree_prompt	44	2	0	42	40	0.02	0.0	38.02
contract_nli_explicit_identification								
neutral_prompt	33	16	72	17	4	0.81		
yesno_prompt	8	8	89	0	12	0.89	0.1096	2.56
agree_prompt	11	11	89	0	9	0.92	0.019	5.5
negated_agree_prompt	97	12	89	0	8	0.93	0.0123	6.26
disagree_prompt	90	7	6	83	13	0.12	0.0	65.98
hearsay								
neutral_prompt	71	38	20	33	3	0.62		
yesno_prompt	64	37	26	27	4	0.67	0.1824	1.78
agree_prompt	62	37	28	25	4	0.69	0.0704	3.27
negated_agree_prompt	35	36	30	23	5	0.7	0.0433	4.08
disagree_prompt	1	0	52	1	41	0.55	0.5557	0.35
contract_nli_confidentiality_of_agreement								
neutral_prompt	81	41	1	40	0	0.51		
yesno_prompt	35	34	40	1	7	0.9	0.0	20.89
agree_prompt	37	36	40	1	5	0.93	0.0	24.75
negated_agree_prompt	45	36	40	1	5	0.93	0.0	24.75
disagree_prompt	36	7	12	29	34	0.23	0.0013	10.3
cuad_non-compete								
neutral_prompt	203	193	211	10	28	0.91		
yesno_prompt	214	200	207	14	21	0.92	0.6892	0.16
agree_prompt	225	211	207	14	10	0.95	0.0216	5.28
negated_agree_prompt	203	215	197	24	6	0.93	0.2684	1.23
disagree_prompt	317	138	42	179	83	0.41	0.0	188.37
agb-de								
neutral_prompt	592	35	161	557	2	0.26		
yesno_prompt	233	19	504	214	18	0.69	0.0	283.4
agree_prompt	155	12	575	143	25	0.78	0.0	344.9
negated_agree_prompt	626	12	601	117	25	0.81	0.0	372.16
disagree_prompt	490	19	247	471	18	0.35	0.0005	12.21
clauses-pl								
neutral_prompt	1583	838	375	745	1495	0.35		
yesno_prompt	1069	580	631	489	1753	0.35	0.977	0.0
agree_prompt	1050	588	658	462	1745	0.36	0.3388	0.92
negated_agree_prompt	1281	1288	236	884	1045	0.44	0.0	83.49
disagree_prompt	2989	2164	295	825	169	0.71	0.0	712.98
overall								
neutral_prompt	4879	2352	1026	2527	1697	0.44		
yesno_prompt	2608	1751	2696	857	2298	0.58	0.0	345.75
agree_prompt	2621	1816	2748	805	2233	0.6	0.0	435.28
negated_agree_prompt	3621	2621	2193	1360	1428	0.63	0.0	650.01
disagree_prompt	6139	3494	908	2645	555	0.58	0.0	285.62

Table 7: Detailed results for the gemma-2-27b-it model

Condition	Pos	TP	TN	FP	FN	Accuracy	P-val	Chi
definition_classification								
neutral_prompt	1279	686	53	593	5	0.55		
yesno_prompt	1253	689	82	564	2	0.58	0.0083	6.96
agree_prompt	1303	690	33	613	1	0.54	0.1058	2.62
negated_agree_prompt	3	690	2	644	1	0.52	0.0	34.69
disagree_prompt	753	522	415	231	169	0.7	0.0	67.85
cuad_cap_on_liability								
neutral_prompt	1178	623	68	555	0	0.55		
yesno_prompt	698	580	505	118	43	0.87	0.0	321.77
agree_prompt	734	606	495	128	17	0.88	0.0	376.76
negated_agree_prompt	433	619	429	194	4	0.84	0.0	347.22
disagree_prompt	682	102	43	580	521	0.12	0.0	471.47
cuad_no-solicit_of_customers								
neutral_prompt	84	42	0	42	0	0.5		
yesno_prompt	41	40	41	1	2	0.96	0.0	33.58
agree_prompt	44	42	40	2	0	0.98	0.0	38.02
negated_agree_prompt	39	42	39	3	0	0.96	0.0	37.03
disagree_prompt	39	0	3	39	42	0.04	0.0	32.09
contract_nli_explicit_identification								
neutral_prompt	63	20	46	43	0	0.61		
yesno_prompt	19	14	84	5	6	0.9	0.0	21.84
agree_prompt	23	17	83	6	3	0.92	0.0	27.22
negated_agree_prompt	86	17	83	6	3	0.92	0.0	27.22
disagree_prompt	80	6	15	74	14	0.19	0.0	31.74
hearsay								
neutral_prompt	56	34	31	22	7	0.69		
yesno_prompt	45	28	36	17	13	0.68	1.0	0.0
agree_prompt	50	33	36	17	8	0.73	0.2207	1.5
negated_agree_prompt	40	34	33	20	7	0.71	0.6171	0.25
disagree_prompt	60	21	14	39	20	0.37	0.0006	11.68
contract_nli_confidentiality_of_agreement								
neutral_prompt	82	41	0	41	0	0.5		
yesno_prompt	41	40	40	1	1	0.98	0.0	35.22
agree_prompt	41	40	40	1	1	0.98	0.0	35.22
negated_agree_prompt	41	40	40	1	1	0.98	0.0	35.22
disagree_prompt	42	2	1	40	39	0.04	0.0	34.22
cuad_non-compete								
neutral_prompt	243	197	175	46	24	0.84		
yesno_prompt	301	213	133	88	8	0.78	0.001	10.78
agree_prompt	362	220	79	142	1	0.68	0.0	43.56
negated_agree_prompt	49	221	49	172	0	0.61	0.0	68.01
disagree_prompt	133	18	106	115	203	0.28	0.0	187.14
agb-de								
neutral_prompt	413	29	334	384	8	0.48		
yesno_prompt	372	26	372	346	11	0.53	0.0028	8.96
agree_prompt	607	33	144	574	4	0.23	0.0	127.71
negated_agree_prompt	417	19	399	319	18	0.55	0.0003	13.32
disagree_prompt	270	7	455	263	30	0.61	0.0	17.95
clauses-pl								
neutral_prompt	3309	2196	7	1113	137	0.64		
yesno_prompt	2537	1560	143	977	773	0.49	0.0	316.8
agree_prompt	2697	1680	103	1017	653	0.52	0.0	279.56
negated_agree_prompt	468	1918	53	1067	415	0.57	0.0	149.89
disagree_prompt	995	848	973	147	1485	0.53	0.0	56.09
overall								
neutral_prompt	6707	3868	714	2839	181	0.6		
yesno_prompt	5307	3190	1436	2117	859	0.61	0.3012	1.07
agree_prompt	5861	3361	1053	2500	688	0.58	0.0	16.68
negated_agree_prompt	1576	3600	1127	2426	449	0.62	0.0001	16.26
disagree_prompt	3054	1526	2025	1528	2523	0.47	0.0	217.89

Table 8: Detailed results for the Llama-3.3-70B model

Condition	Pos	TP	TN	FP	FN	Accuracy	P-val	Chi
definition_classification								
neutral_prompt	562	520	604	42	171	0.84		
yesno_prompt	734	658	570	76	33	0.92	0.0	46.13
agree_prompt	758	661	549	97	30	0.91	0.0	29.13
negated_agree_prompt	528	671	508	138	20	0.88	0.0014	10.16
disagree_prompt	1316	681	11	635	10	0.52	0.0	242.51
cuad_cap_on_liability								
neutral_prompt	828	619	414	209	4	0.83		
yesno_prompt	209	202	616	7	421	0.66	0.0	73.98
agree_prompt	215	210	618	5	413	0.66	0.0	67.89
negated_agree_prompt	969	268	614	9	355	0.71	0.0	40.83
disagree_prompt	1103	612	132	491	11	0.6	0.0	265.0
cuad_no-solicit_of_customers								
neutral_prompt	84	42	0	42	0	0.5		
yesno_prompt	40	40	42	0	2	0.98	0.0	34.57
agree_prompt	40	40	42	0	2	0.98	0.0	34.57
negated_agree_prompt	43	41	42	0	1	0.99	0.0	37.21
disagree_prompt	83	42	1	41	0	0.51	1.0	0.0
contract_nli_explicit_identification								
neutral_prompt	14	9	84	5	11	0.85		
yesno_prompt	9	8	88	1	12	0.88	0.5791	0.31
agree_prompt	13	11	87	2	9	0.9	0.3017	1.07
negated_agree_prompt	95	12	87	2	8	0.91	0.1814	1.79
disagree_prompt	96	10	3	86	10	0.12	0.0	67.84
hearsay								
neutral_prompt	31	25	47	6	16	0.77		
yesno_prompt	28	25	50	3	16	0.8	0.505	0.44
agree_prompt	30	26	49	4	15	0.8	0.3711	0.8
negated_agree_prompt	69	21	49	4	20	0.74	0.7237	0.12
disagree_prompt	68	19	4	49	22	0.24	0.0	27.76
contract_nli_confidentiality_of_agreement								
neutral_prompt	44	38	35	6	3	0.89		
yesno_prompt	38	37	40	1	4	0.94	0.3428	0.9
agree_prompt	40	39	40	1	2	0.96	0.1138	2.5
negated_agree_prompt	43	38	40	1	3	0.95	0.1824	1.78
disagree_prompt	48	7	0	41	34	0.09	0.0	62.13
cuad_non-compete								
neutral_prompt	205	196	212	9	25	0.92		
yesno_prompt	205	195	211	10	26	0.92	0.8312	0.05
agree_prompt	193	183	211	10	38	0.89	0.0056	7.68
negated_agree_prompt	228	201	208	13	20	0.93	1.0	0.0
disagree_prompt	411	221	31	190	0	0.57	0.0	116.63
agb-de								
neutral_prompt	143	14	589	129	23	0.8		
yesno_prompt	159	8	567	151	29	0.76	0.0254	4.99
agree_prompt	331	17	404	314	20	0.56	0.0	147.57
negated_agree_prompt	297	25	285	433	12	0.41	0.0	260.75
disagree_prompt	680	34	72	646	3	0.14	0.0	420.54
clauses-pl								
neutral_prompt	2465	1448	103	1017	885	0.45		
yesno_prompt	1668	932	384	736	1401	0.38	0.0	53.63
agree_prompt	2452	1461	129	991	872	0.46	0.1086	2.57
negated_agree_prompt	875	1600	142	978	733	0.5	0.0	58.89
disagree_prompt	3145	2236	211	909	97	0.71	0.0	633.72
overall								
neutral_prompt	4376	2911	2088	1465	1138	0.66		
yesno_prompt	3090	2105	2568	985	1944	0.61	0.0	49.96
agree_prompt	4072	2648	2129	1424	1401	0.63	0.0	28.07
negated_agree_prompt	3147	2877	1975	1578	1172	0.64	0.0007	11.39
disagree_prompt	6950	3862	465	3088	187	0.57	0.0	133.29

Table 9: Detailed results for the gpt-4o-2024-08-06 model

Condition	Pos	TP	TN	FP	FN	Accuracy	P-val	Chi
definition_classification								
neutral_prompt	508	449	587	59	242	0.77		
yesno_prompt	711	630	565	81	61	0.89	0.0	81.85
agree_prompt	804	663	505	141	28	0.87	0.0	47.41
negated_agree_prompt	422	682	413	233	9	0.82	0.0062	7.49
disagree_prompt	525	130	251	395	561	0.28	0.0	535.31
cuad_cap_on_liability								
neutral_prompt	1226	623	20	603	0	0.52		
yesno_prompt	214	209	618	5	414	0.66	0.0	33.09
agree_prompt	272	265	616	7	358	0.71	0.0	58.88
negated_agree_prompt	926	305	608	15	318	0.73	0.0	79.87
disagree_prompt	910	339	52	571	284	0.31	0.0	177.97
cuad_no-solicit_of_customers								
neutral_prompt	84	42	0	42	0	0.5		
yesno_prompt	38	38	42	0	4	0.95	0.0	29.76
agree_prompt	35	35	42	0	7	0.92	0.0	23.59
negated_agree_prompt	42	42	42	0	0	1.0	0.0	40.02
disagree_prompt	44	3	1	41	39	0.05	0.0	34.22
contract_nli_explicit_identification								
neutral_prompt	32	15	72	17	5	0.8		
yesno_prompt	17	15	87	2	5	0.94	0.0007	11.53
agree_prompt	20	16	85	4	4	0.93	0.0005	12.07
negated_agree_prompt	87	16	83	6	4	0.91	0.0033	8.64
disagree_prompt	87	3	5	84	17	0.07	0.0	64.04
hearsay								
neutral_prompt	28	24	49	4	17	0.78		
yesno_prompt	39	31	45	8	10	0.81	0.5791	0.31
agree_prompt	38	30	45	8	11	0.8	0.7518	0.1
negated_agree_prompt	53	31	43	10	10	0.79	1.0	0.0
disagree_prompt	55	21	19	34	20	0.43	0.0001	15.28
contract_nli_confidentiality_of_agreement								
neutral_prompt	80	41	2	39	0	0.52		
yesno_prompt	35	35	41	0	6	0.93	0.0	22.76
agree_prompt	38	37	40	1	4	0.94	0.0	25.93
negated_agree_prompt	38	38	35	6	3	0.89	0.0	23.36
disagree_prompt	41	5	5	36	36	0.12	0.0	23.81
cuad_non-compete								
neutral_prompt	171	167	217	4	54	0.87		
yesno_prompt	195	184	210	11	37	0.89	0.1227	2.38
agree_prompt	210	196	207	14	25	0.91	0.0073	7.2
negated_agree_prompt	222	203	204	17	18	0.92	0.0017	9.88
disagree_prompt	256	83	48	173	138	0.3	0.0	208.21
agb-de								
neutral_prompt	132	13	599	119	24	0.81		
yesno_prompt	283	19	454	264	18	0.63	0.0	76.48
agree_prompt	234	16	500	218	21	0.68	0.0	40.65
negated_agree_prompt	550	18	531	187	19	0.73	0.0	22.75
disagree_prompt	520	18	216	502	19	0.31	0.0	272.28
clauses-pl								
neutral_prompt	1499	779	400	720	1554	0.34		
yesno_prompt	1082	573	611	509	1760	0.34	0.899	0.02
agree_prompt	1955	1219	384	736	1114	0.46	0.0	176.11
negated_agree_prompt	1293	1379	339	781	954	0.5	0.0	247.6
disagree_prompt	2588	1861	393	727	472	0.65	0.0	506.13
overall								
neutral_prompt	3760	2153	1946	1607	1896	0.54		
yesno_prompt	2614	1734	2673	880	2315	0.58	0.0	34.73
agree_prompt	3606	2477	2424	1129	1572	0.64	0.0	236.4
negated_agree_prompt	3633	2714	2298	1255	1335	0.66	0.0	292.15
disagree_prompt	5026	2463	990	2563	1586	0.45	0.0	92.37