

Terms and Conditions Apply: Evaluating Consumer LLM Reliability in NDA Interpretation to Investigate the Non-Linear Role of Reasoning.

Shaurya Jain Mathur*

Independent Researcher
shauryajmathur@gmail.com

Simeon Sayer

Harvard University, Cambridge, MA, USA
ssayer@fas.harvard.edu

*Primary author

Abstract

This paper evaluates how advances in general-purpose large language models affect routine consumer legal interpretation, **using NDAs as a case study**, and what aggregate benchmark gains can conceal. Gemini 2.5 Flash-Lite and GPT-5.5 classify 2,091 NDA-hypothesis pairs from ContractNLI as Entailment, Contradiction, or NotMentioned. GPT performs better overall, especially on contradictions, but Gemini is uniquely correct on 247 instances and more often preserves NotMentioned when contracts are silent. GPT more readily infers unsupported legal relationships and sometimes converts absence into false contradiction. GPT alone reaches 80.2% accuracy, while a descriptive perfect-selector ceiling that chooses a correct system whenever either is correct reaches 92.0%. These results show that higher aggregate performance can reflect an exchange of capabilities rather than uniform improvement. For **consumer-facing NDA interpretation**, the models remain useful for preliminary contract review but are not reliable enough to act as autonomous legal advisers. More broadly, the study argues that evaluations of legal LLMs should report error direction, correctness overlap, and task-level reversals alongside aggregate metrics so that model progress is judged by which behaviours improve, which regress, and how those changes affect practical reliability.

1 Introduction

The emerging role of Artificial Intelligence (AI) in the courtroom and beyond are widely discussed. Developments in capabilities and broadening use cases raise familiar questions as to whether AI can be trusted to perform law-related tasks and how far it should replace human judgment. However, beyond the legal professionals, a broader change may already be under way.

The average person will not first encounter legal AI through their role as a judge, or a specialised

legal profession platform. It is much more likely that they ask a general-purpose model, like Gemini or ChatGPT, a routine question about a standard legal text, such as a mortgage, a prenup, or in this case what an NDA permits them to share. The potential value is intuitive: it saves checking with a lawyer, and reduces risk of non-compliance by misinterpretation. Provided, of course, on there being a real basis for trusting the answer the model gives.

This exposes a practical question: can widely used large language models reliably support routine legal interpretation, or do their errors make them unsafe for real-world use? Answering that question properly requires analysis that goes beyond how often a model is correct, but that explores what it does when it is wrong.

Existing research has begun to evaluate artificial intelligence across a wide range of legal tasks. LegalBench and LawBench test large language models across multiple forms of legal reasoning, while frameworks such as HELM argue for standardised, multi-metric evaluation rather than reliance on a single aggregate score (Guha et al., 2023; Fei et al., 2024; Bommasani et al., 2023). Aggregate scores are useful: they show that one model gets more answers right. What they do not show, on their own, is whether those gains are uniform, whether the stronger model loses cases the weaker model once answered correctly, or whether one failure mode has simply been exchanged for another.

Legal AI does not merely need to retrieve legal information. It must interpret dense language, apply rules to the available context, and preserve the distinction between what is implied and what is absent, all while avoiding unsupported claims (Dahl et al., 2024; Magesh et al., 2025). That distinction has practical consequences. A false permission, an omitted obligation, and an invented restriction may each count as one error, but they can direct an un-

informed user towards entirely different decisions.

This paper examines that evaluation problem through a practical legal task: whether consumer-accessible large language models can accurately interpret non-disclosure agreements. NDAs are useful for two reasons. They are common enough to be encountered by ordinary people and small organisations, yet standardised enough for the same legal propositions to be tested repeatedly. ContractNLI formalises this structure through 17 fixed hypotheses and three labels, Entailment, Contradiction, and NotMentioned, across 123 test documents.

The study compares two general-purpose model systems from families accessible to ordinary users, although the experimental runs were API-mediated rather than conducted through consumer chat interfaces. Gemini 2.5 Flash-Lite represents the lighter, speed-oriented end of the comparison; GPT-5.5 represents the higher-capability end. The purpose is not merely to rank them. It is to test whether greater aggregate capability produces balanced improvement, or whether it changes the kinds of errors the model makes.

The results show why that distinction matters. GPT achieved the stronger aggregate score and was far more successful at identifying contradictions. Gemini, however, was uniquely correct on 247 instances and outperformed GPT on several legal propositions, often because it was more willing to preserve NotMentioned when the contract was silent. GPT’s greater willingness to infer a legal relationship also produced more false contradictions. A hypothetical selector that always chose the correct system would reach 92.0% accuracy, although no such selector could be deployed without already knowing the gold label.

The significance of the study therefore extends beyond consumer contract review. High overall scores can conceal structural failures, default response tendencies, and capabilities that disappear when one model replaces another. By combining confusion matrices, hypothesis-level comparisons, and correctness overlap, this paper asks not only whether a benchmark score rises, but what changed underneath it, what was lost, and whether the resulting pattern of errors is sufficiently manageable for consumers to rely on an LLM when reading **an NDA**.

2 Background

Legal natural language processing research has progressively tried to measure whether computational models can perform tasks that require legal interpretation. LexGLUE standardised evaluation across several English-language legal understanding tasks, while LawBench and LegalBench expanded this approach across a wider range of legal questions and levels of reasoning (Chalkidis et al., 2022; Guha et al., 2023; Fei et al., 2024).

These benchmarks helped establish that legal capability is not one skill, but a combination of information retrieval, rule application, interpretation, and inference. Their broad scores, however, can conceal substantial variation across the capabilities being tested. General model evaluation has reached a similar conclusion: held-out accuracy can overstate practical reliability when systematic behavioural weaknesses remain hidden beneath the average (Ribeiro et al., 2020; Bommasani et al., 2023).

Contract analysis provides a particularly useful setting in which to examine these weaknesses because the task is comparatively objective and the relevant skills recur across documents. CUAD introduced expert-annotated data for identifying important contractual clauses. ContractNLI similarly frames contract review as a three-way determination of whether a legal proposition is entailed, contradicted, or not mentioned. The original ContractNLI study found that existing models struggled with long documents and constructions such as negation by exception (Hendrycks et al., 2021; Koreeda and Manning, 2021). Later work showed that prompting can improve predictive performance, but has focused more heavily on raising the score than on explaining the practical meaning and distribution of the remaining errors (Yao et al., 2024).

Taken together, this literature shows why a high score is informative but incomplete. ContractNLI is therefore used here not only to benchmark two general-purpose systems, but to examine how their errors move across labels, legal propositions, and documents. The central question is not simply which score is higher. It is what changed beneath the difference.

3 Dataset

NDAs provide a useful setting for this analysis because they are both socially familiar and unusually

standardised. They do not represent the full range of legal reasoning, but they allow the same propositions and the same directions of error to be tested repeatedly.

The primary dataset is ContractNLI, which was designed for document-level natural language inference on contracts (Koreeda and Manning, 2021). It contains 423 training NDAs, 61 development NDAs, and 123 test NDAs. This study uses the complete test split.

ContractNLI uses 17 fixed legal hypotheses developed by its authors and reviewed by paralegals. Across the 123 test documents, this produces exactly 2,091 NDA-hypothesis classification instances. Each instance receives one gold label: Entailment, Contradiction, or NotMentioned.

A computational linguistics researcher served as the primary annotator, assisted by two selected Amazon Mechanical Turk workers; the primary annotator then merged, reviewed, and adjusted the annotations (Koreeda and Manning, 2021). The NDAs were collected through internet search engines and the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system before being screened for relevance by Koreeda and Manning.

The corpus is restricted to unilateral or bilateral NDAs and confidentiality agreements between two parties. Employer-employee NDAs are excluded.

This restriction makes the dataset more controlled: each document belongs to roughly the same legal category and can be tested against roughly the same set of propositions. It also imposes an important limit. The study is not an evaluation of LLMs across all legal documents, and its findings should not be generalised automatically to other forms of legal text. It measures performance on a narrower but still consequential object: public English-language NDAs, which are common, relatively standardised, and often encountered by individuals or small organisations without easy access to professional legal advice.

No training or fine-tuning was conducted on the ContractNLI training set. The training and development splits were used only to understand the dataset structure and label format; all evaluation was conducted on the test split. This follows directly from the research question. The aim is not to build a specialised legal model, but to test how well accessible general-purpose systems perform without task-specific training, which is closer to the experience of a layperson consulting an ordinary AI tool about a contract.

The analysis therefore uses three parts of ContractNLI: the NDA text, the 17 hypotheses, and the curated gold labels. The document supplies the context, the hypothesis supplies the legal claim to be assessed, and the gold label provides the basis for evaluating each model output.

This three-way structure makes it possible to measure not only whether a prediction is wrong, but how it is wrong. Predicting NotMentioned when a contract contradicts a proposition reflects a different failure from inventing Entailment, even though both count as one error under an accuracy metric. Repeating the 17 hypotheses across documents therefore turns them into standardised probes of model behaviour.

4 Methodology

4.1 Research Design

This study uses a comparative evaluation design to test whether general-purpose large language models can interpret non-disclosure agreements reliably. It does not train a new legal model or optimise a system specifically for ContractNLI. Instead, it evaluates the broadly accessible model families that an ordinary user might consult when trying to understand a contract without legal assistance.

The task is framed as document-level natural language inference. For each test instance, the system receives the complete NDA and one of the 17 fixed legal hypotheses, then classifies the relationship between them:

- Entailment: the contents of the contract support the hypothesis.
- Contradiction: the contract establishes a rule incompatible with the hypothesis.
- NotMentioned: the contract neither supports nor contradicts the hypothesis.

No training, fine-tuning, or parameter updating was carried out. The training and development splits were used only to understand the task structure, label definitions, and hypothesis format. Every reported result is calculated on the complete 123-document test set.

4.2 Evaluated Systems

The first evaluated system was Gemini 2.5 Flash-Lite, accessed through Google’s Gemini API under the exact model identifier `gemini-2.5-flash-lite`. Temperature was set to 0.0 to reduce sampling

variation, although this does not guarantee deterministic output. The model received no examples, chain-of-thought, evidence spans, or additional legal background. Each NDA-hypothesis pair was submitted in a separate request so that earlier outputs could not influence later predictions. The complete run produced 2,091 responses, with an average response time of approximately 0.731 seconds per instance.

As each system was evaluated using a single complete run should therefore be interpreted as single-run estimates rather than measurements of run-to-run stability. A repeated-run stability analysis was not conducted due to the additional computational and API cost it would require.

Gemini 2.5 Flash-Lite represents the lighter, speed-oriented end of the general-purpose model spectrum. Its inclusion tests whether a system designed for rapid and inexpensive inference can still provide reliable legal interpretation at scale.

The second evaluated system was GPT-5.5, accessed through the OpenAI API under the exact model identifier gpt-5.5. As with Gemini, the model received no examples, chain-of-thought, evidence spans, or additional legal background, and each NDA-hypothesis pair produced one independent response.

GPT-5.5 represents the more advanced end of the general-purpose model spectrum. The comparison is therefore designed not simply to produce a ranking, but to test whether greater model capability is associated with more reliable and balanced legal interpretation, and whether any improvement changes the errors that remain.

Both systems are proprietary, and their training data, model weights, and internal computational processes were unavailable for inspection. The comparison is therefore behavioural. It also does not isolate reasoning as a causal treatment, because provider, training, architecture, context handling, and inference design all change at once. References below to 'reasoning', 'overthinking', or inferential assertiveness describe observable output patterns and hypotheses for matched-model experiments; they are not direct observations of internal computation.

Both models were chosen deliberately to represent opposite ends of the consumer-model spectrum: a flagship and a light-weight efficiency-oriented model. The comparison is therefore not intended as a like-for-like assessment of providers, but to examine how systems

with substantially different capability–cost profiles perform on the same NDA interpretation task

4.3 Prompt and Inference Procedure

Every instance used the same prompt structure:

Listing 1: Prompt used for every NDA-hypothesis pair

```
"You are a contract analysis assistant. "
f"Classify whether the statement '{
    row_hypothesis}' is "
f"'Entailment', 'Contradiction', or '
    NotMentioned' "
f"in the following NDA document. "
f"Return ONLY the single word label.\n\n"
f"Document:\n{text}\n"
```

The design served two purposes.

1. First, each attempt remained independent, preventing an earlier interpretation from influencing a later answer.
2. Second, the prompt approximated a plausible consumer interaction with a general-purpose model about one focused contractual question, rather than asking for a complete benchmark annotation in a single request.

The prompt was deliberately minimal. The label names were supplied, but Entailment, Contradiction, and NotMentioned were not separately defined, and the model received no legal context beyond the NDA itself. This reduced prompt engineering as a variable and made comprehension of the concise instructions part of the measured task. The label-only output also standardised scoring by preventing explanation length or writing quality from affecting the result. For each of the 123 NDAs, the evaluation script retrieved the relevant hypothesis, inserted the complete document and hypothesis into the fixed prompt, submitted the request, and recorded the response, gold label, match status, and response time. A Python script then normalised the output by taking the first word and stripping surrounding whitespace.

Each normalised output was compared case-insensitively with the corresponding ContractNLI gold label. The final audit contained all 2,091 expected rows, no duplicate contract-hypothesis keys, and matching gold labels across the two model runs.

The minimal prompt reduces one source of experimental discretion, but it also means that understanding the label scheme is part of the task. The

results should therefore be read as end-to-end performance under concise instructions, not as legal reasoning measured in complete isolation.

4.4 Evaluation Metrics and Diagnostic Framework

The evaluation reports aggregate accuracy, macro F1, weighted F1, and balanced accuracy. Because Contradiction accounts for only 10.52% of the test set, class-sensitive metrics are included to prevent majority-label performance from dominating the comparison. Full formulae and statistical details appear in Appendix B.

Even these metrics can conceal systematic behaviour when they are read only in aggregate. The analysis therefore operates at several levels. Overall and hypothesis-specific confusion matrices show the direction of error; incorrect predictions are grouped into hallucinated provisions, omitted provisions, and polarity reversals; and every instance is assigned to one of four overlap categories: both systems correct, GPT only correct, Gemini only correct, or both systems wrong.

The overlap counts also support a descriptive perfect-selector ceiling: the maximum accuracy of a hypothetical system that always chooses a correct model whenever either model is right. This is not a deployable ensemble, since it assumes access to the gold label. Its purpose is narrower. It shows how much correct behaviour disappears when the comparison is reduced to a single winning system.

Together, these diagnostics allow the study to ask three related questions: which system scores higher, what changed underneath that score, and what a more selective system might preserve.

5 Results and Discussion

Appendix A provides the full hypothesis-level results, confusion matrices, and detailed clause-group analysis.

5.1 Overall Performance and Correctness Overlap

GPT’s aggregate advantage is clear. It achieved 80.20% accuracy, a macro F1 of 0.757, and higher weighted F1 and balanced accuracy than Gemini, which achieved 75.27% accuracy and a macro F1 of 0.625. On a conventional leaderboard, the 4.93-point accuracy gap would be the principal result. It is not, however, the whole comparison.

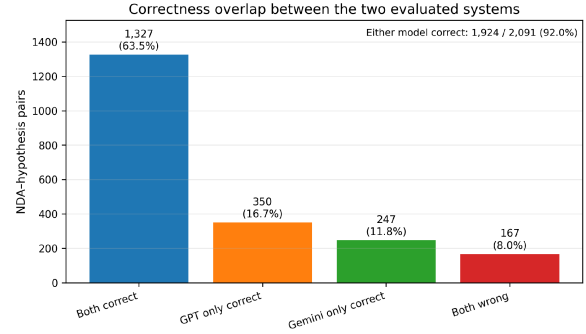


Figure 1: Correctness overlap between GPT-5.5 and Gemini 2.5 Flash-Lite. The 92.0% value is a descriptive perfect-selector ceiling, not a deployable ensemble result.

The overlap reveals 350 cases in which only GPT was correct and 247 in which only Gemini was correct. A hypothetical selector that chose the correct system whenever either succeeded would therefore classify 1,924 of 2,091 instances correctly, reaching 92.0% accuracy. This does not show that those cases can be routed in practice. It does show that the lower-scoring model retained a substantial body of correct behaviour that the aggregate winner did not absorb.

Both systems also perform meaningfully above the 67.38% per-hypothesis majority baseline. Yet replacing Gemini with GPT adds 350 correct answers while losing 247 that Gemini got right. Reporting only 75.3% and 80.2% captures the net gain, but not the answers, behaviours, or possible capabilities that moved in the opposite direction.

5.2 Confusion Matrices and Error Profiles

The overall confusion matrices show that the two scores were produced by different error profiles, not by a uniform reduction in every kind of mistake.

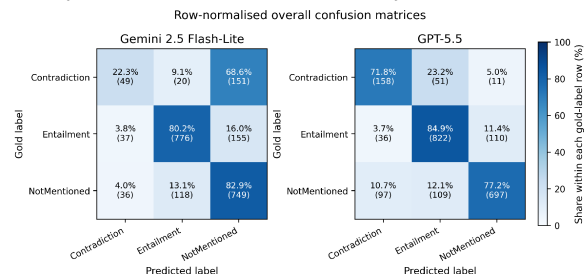


Figure 2: Row-normalised overall confusion matrices for Gemini 2.5 Flash-Lite and GPT-5.5. Percentages are calculated within each gold-label row; counts appear in parentheses.

Gemini’s dominant error was to classify 151 of the 220 gold Contradiction cases as NotMentioned, leaving Contradiction recall at 22.3%. GPT

System	Correct	Accuracy	Macro F1	Weighted F1	Balanced accuracy
GPT-5.5	1,677 / 2,091	80.20%	0.757	0.805	0.780
Gemini 2.5 Flash-Lite	1,574 / 2,091	75.27%	0.625	0.742	0.618
Difference	103	+4.93 pp	+0.132	+0.063	+0.162

Table 1: Aggregate performance of the two evaluated systems.

made that error only 11 times and recalled 71.8% of contradictions. In behavioural terms, the more advanced system was far more willing to reject a proposition when the contract supplied evidence against it.

That willingness also produced a different failure. GPT classified 97 gold NotMentioned cases as Contradiction, compared with 36 for Gemini, and its NotMentioned recall fell from 82.9% to 77.2%. Gemini therefore exhibits an omission-dominant profile; GPT exhibits a more assertive inferential profile. The pattern is consistent with the possibility that further reasoning is especially useful for rejecting claims, but can sometimes cause a model to overinterpret contractual silence and reason itself into a contradiction. Since internal reasoning was not observed, this remains a behavioural interpretation rather than an established mechanism.

The higher score, then, came with a trade-off. More assertive inference helped GPT push back against unsupported claims, but the same tendency sometimes transformed silence into an invented restriction. Which profile is preferable depends on the cost attached to each error, and this study does not impose a universal legal cost model.

The error taxonomy makes the same pattern visible in another form. Nearly half of GPT’s errors were hallucinated provisions, while 59.2% of Gemini’s were omitted provisions (see **Appendix B for definitions and additional diagnostics**). Inventing a permission or restriction is not equivalent to overlooking a real obligation, even if both reduce accuracy by one. Without an application-specific cost model, neither profile can be treated as categorically safer.

5.3 Hypothesis- and Contract-Level Differences

Performance also varied substantially across the 17 legal hypotheses and the 123 individual NDAs. If GPT simply dominated Gemini, the aggregate advantage would recur consistently across both levels. It did not.

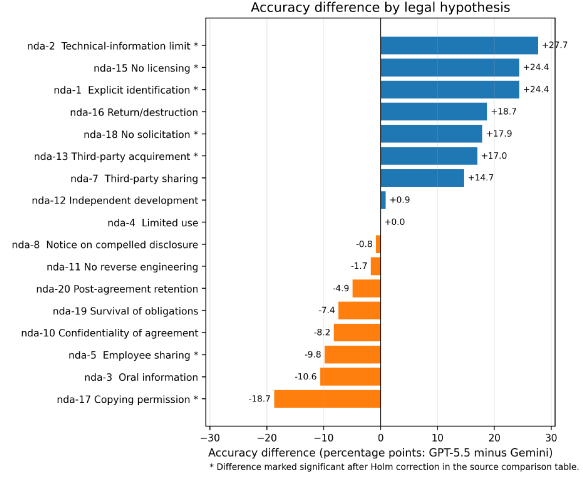


Figure 3: Accuracy difference by legal hypothesis (GPT-5.5 minus Gemini 2.5 Flash-Lite). Positive values favour GPT; negative values favour Gemini. An asterisk marks differences identified as significant after Holm correction in the source comparison table.

Of the 17 hypotheses, eight favoured GPT, eight favoured Gemini, and one was tied. The source comparison table marks seven differences as statistically significant after Holm correction. GPT’s largest advantages concerned technical-information limits, explicit identification, no licensing, no solicitation, and return or destruction. Gemini’s largest advantage concerned copying permission, with further gains on oral information, employee sharing, confidentiality of the agreement, and survival of obligations.

Those reversals become more legible when the full confusion matrices are examined. Two hypotheses provide a useful pair because each rewards one model’s broader response tendency and punishes the other.

Error category	Definition	GPT-5.5	Gemini 2.5 Flash-Lite
Hallucinated provision	Gold NotMentioned → Entailment or Contradiction	206 (49.8%)	154 (29.8%)
Omitted provision	Gold Entailment or Contradiction → NotMentioned	121 (29.2%)	306 (59.2%)
Polarity reversal	Entailment ↔ Contradiction	87 (21.0%)	57 (11.0%)

Table 2: Legally meaningful error categories and their share of each system’s errors.

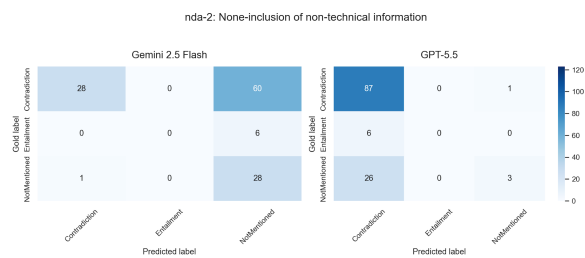


Figure 4: Hypothesis-specific confusion matrices for nda-2 (technical-information limit). GPT’s contradiction-heavy response profile is rewarded by the local label distribution, while Gemini more often preserves NotMentioned.

On nda-2, GPT achieved 73.2% accuracy against Gemini’s 45.5%, but the matrix shows why that gap should not be read as a general understanding advantage. Neither system identified any of the six Entailment cases. GPT instead predicted Contradiction in 119 of 123 instances, correctly capturing 87 of 88 gold Contradictions while recognising only 3 of 29 NotMentioned cases. Gemini did almost the reverse: it recognised 28 of 29 NotMentioned cases but only 28 of 88 Contradictions. The score gap was therefore produced largely by which default aligned with the local label distribution.

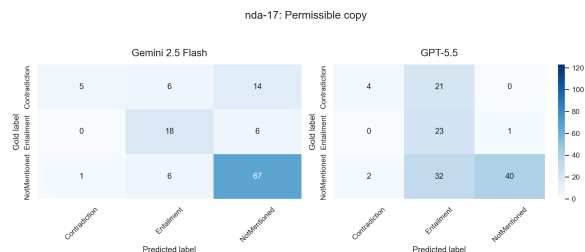


Figure 5: Hypothesis-specific confusion matrices for nda-17 (copying permission). Here Gemini’s tendency to preserve NotMentioned becomes an advantage, while GPT overextends references to copies into permission to make them.

nda-17 reverses the ranking. Gemini achieved 73.2% accuracy and GPT 54.5%. GPT correctly identified 23 of 24 Entailments, but it also classified 21 of 25 Contradictions and 32 of 74 NotMentioned cases as Entailment. Gemini preserved 67 of the 74 NotMentioned cases and therefore won the aggregate comparison, even though it classified 14 of the 25 Contradictions as NotMentioned. The same caution that harmed Gemini on nda-2 helped it here; the same inferential assertiveness that helped GPT reject propositions elsewhere caused it to infer a permission the benchmark did not support.

Taken together, the two matrices provide a concrete version of the paper’s broader claim. Model improvement did not move every capability in one direction. It changed the balance between recognising a legal relationship and preserving contractual silence. Which system appears better therefore depends not only on the legal proposition, but on how often each label occurs and which error tendency that proposition rewards.

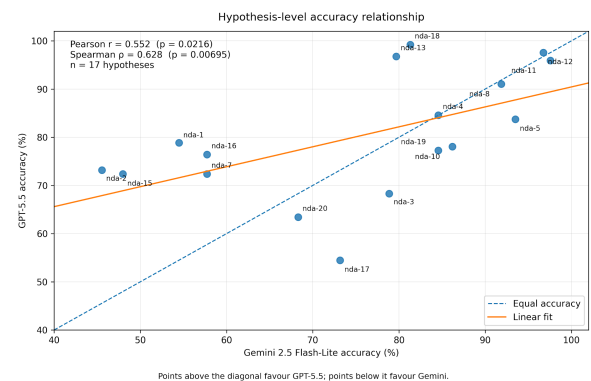


Figure 6: Hypothesis-level accuracy comparison. Points above the diagonal favour GPT-5.5; points below it favour Gemini. The moderate positive correlation shows some shared notion of difficulty, while the dispersion around the diagonal shows that neither system uniformly subsumes the other.

Figure 6 extends the two examples across the full hypothesis set. The moderate Pearson ($r = 0.552$) and Spearman ($\rho = 0.628$) correlations show that the systems often agreed about which propositions were broadly easy or difficult. Yet the points remain widely dispersed around the equal-performance line: *nda-2* and *nda-1* sit well above it, while *nda-17* and *nda-3* sit below it. The benchmark therefore contains both shared difficulty and model-specific reversals. The headline score records the average of those movements, but not their direction.

At the document level, the relationship was weaker still. Contract-level accuracies were almost uncorrelated (Pearson $r = 0.074$; Spearman $\rho = 0.090$), whereas hypothesis-level accuracies showed the moderate relationships plotted above ($r = 0.552$; $\rho = 0.628$). The systems therefore agreed more on broad proposition difficulty than on which particular NDA-hypothesis combination would defeat them. The remaining variation lies in the interaction between the proposition, its expression in a particular contract, and each model’s response threshold.

5.4 Conditions That Change the Model Comparison

The ranking also depends on which documents and legal constructions the benchmark contains.

5.4.1 Contract Length

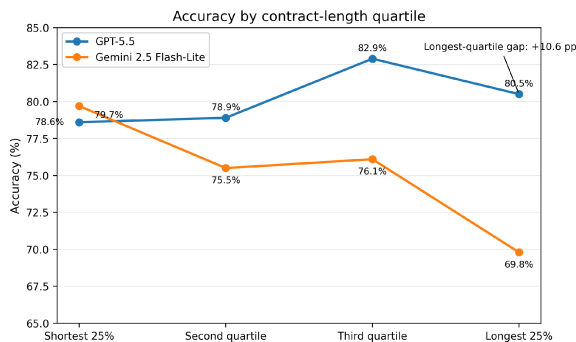


Figure 7: Accuracy by contract-length quartile. The aggregate advantage of GPT-5.5 emerged progressively as the documents became longer.

Gemini slightly outperformed GPT on the shortest contracts, while GPT’s advantage exceeded ten percentage points in the longest quartile. The relationship is correlational: longer NDAs may also differ in drafting style, clause density, source, or legal complexity, so a controlled experiment would be needed to isolate length itself.

5.4.2 Negation by Exception

Across seven exception-associated hypotheses, both systems became less reliable. GPT’s accuracy fell to 73.52%, essentially eliminating its aggregate accuracy advantage over Gemini at 73.75%, although GPT retained the higher macro F1 (0.672 versus 0.607). This does not show that Gemini generally handles exceptions better; it shows that benchmark composition can materially change the headline comparison. Full subset results appear in Appendix C.

5.5 Hypotheses and Future Scope

The error patterns suggest four mechanisms for controlled follow-up experiments: quantifier and scope reasoning, permission expressed as an exception to a prohibition, negative constructions, and temporal or conditional triggers. In each case a model can locate the relevant clause yet still apply its legal force too broadly, too narrowly, or at the wrong time.

The broader implication is that useful legal reasoning is not simply “more reasoning” on every instance. A stronger system should infer further when the document contains qualified evidence while preserving NotMentioned when the relevant provision is absent. Matched reasoning-on/off variants, evidence-span requirements, calibrated abstention, and learned routing are direct next tests. Appendix D gives the fuller mechanism-by-mechanism hypotheses and experimental directions.

In the future, the scope could extend the evaluation to open-weight models, enabling further comparison across model sizes, prompting and inference configuration. This would also improve reproducibility and permit closer investigation of whether the observed error profiles arise across model families.

6 Conclusion

Across 2,091 NDA-hypothesis pairs, GPT-5.5 achieved higher overall accuracy and macro F1 than Gemini 2.5 Flash-Lite, especially by recognising contradictions more effectively. The improvement was not uniform: Gemini was uniquely correct on 247 instances and outperformed GPT on several hypotheses, often because it more readily preserved NotMentioned when the contract was silent. GPT’s greater willingness to infer legal relationships also produced false contradictions. Nei-

ther system was consistently reliable across the full task.

The central benchmarking result is therefore not simply that one model scored 4.9 percentage points higher. Aggregate accuracy concealed reversals in which the nominally weaker system retained useful behaviour that the stronger one lost. Confusion matrices, hypothesis-level comparisons, correctness overlap, and condition-specific analyses expose those exchanges. The 92.0% perfect-selector ceiling is not deployable, but it quantifies how much correct behaviour remains outside a single winning model and motivates selective routing, abstention, and evidence-grounded verification.

For consumers **interpreting NDAs**, these systems can assist with preliminary issue spotting and clause discovery, but the results do not support treating them as autonomous legal advisers. More generally, progress in legal LLM evaluation should be judged not only by whether a benchmark score rises, but by which errors disappear, which new errors appear, and whether those changes make model behaviour safer and more predictable for the task at hand.

Limitations

First, the study compares two different proprietary systems rather than one model with reasoning toggled. Their output profiles are consistent with a trade-off involving reasoning or inferential assertiveness, but provider, training data, architecture, context handling, and calibration are all confounded. Any causal claim would require matched ablations.

Second, the evaluated model families are relevant to consumers, but the experiment was API-mediated. Consumer interfaces may add system prompts, safety layers, retrieval, or different model snapshots. The minimal label-only prompt also makes task-format inference part of the evaluation and may disadvantage a system that would improve under explicit label definitions.

Third, ContractNLI is a curated gold standard for public English-language NDAs, not a measure of every legal document or of personalised legal advice. A temperature setting of 0.0 does not guarantee deterministic output, the study reports one complete run per system, and the perfect-selector ceiling is descriptive rather than deployable. Nor does the analysis impose a universal cost model on hallucinated, omitted, and polarity-reversed provi-

sions; the practical risk of each error depends on the user and application.

References

- Rishi Bommasani, Percy Liang, and Tony Lee. 2023. [Holistic Evaluation of Language Models](#). *Annals of the New York Academy of Sciences*, 1525(1):140–146.
- Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. [LexGLUE: A Benchmark Dataset for Legal Language Understanding in English](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.
- Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. [Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models](#). *Journal of Legal Analysis*, 16(1):64–93.
- Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall, New York.
- 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 E. Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. [LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models](#). *Advances in Neural Information Processing Systems*, 36, Datasets and Benchmarks Track.
- Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. [CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Sture Holm. 1979. [A Simple Sequentially Rejective Multiple Test Procedure](#). *Scandinavian Journal of Statistics*, 6(2):65–70.

- Yuta Koreeda and Christopher Manning. 2021. [ContractNLI: A Dataset for Document-Level Natural Language Inference for Contracts](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2025. [Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools](#). *Journal of Empirical Legal Studies*, 22(2):216–242.
- Savinay Narendra, Kaushal Shetty, and Adwait Ratnaparkhi. 2024. [Enhancing Contract Negotiations with LLM-Based Legal Document Comparison](#). In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 143–153, Miami, FL, USA. Association for Computational Linguistics.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. [Beyond Accuracy: Behavioral Testing of NLP Models with CheckList](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.
- Bingsheng Yao, Guiming Chen, Ruishi Zou, Yuxuan Lu, Jiachen Li, Shao Zhang, Yisi Sang, Sijia Liu, James Hendler, and Dakuo Wang. 2024. [More Samples or More Prompts? Exploring Effective Few-Shot In-Context Learning for LLMs with In-Context Sampling](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1772–1790, Mexico City, Mexico. Association for Computational Linguistics.

A Full Hypothesis-Level Results

This appendix preserves the complete proposition-level evidence while allowing the main discussion to remain focused. Figure 3 provides the summary used in the body of the paper; Figures 4 and 5 reproduce nda-2 and nda-17 as illustrative reversals. The table, complete matrix set, and clause-group discussion below report the full result for each hypothesis.

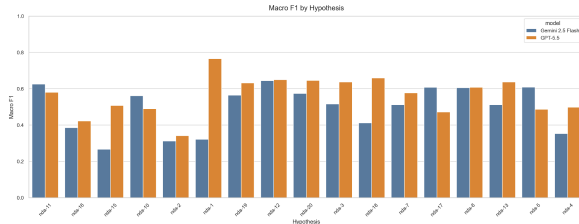


Figure 8: Three-class macro F1 by hypothesis. Values for hypotheses with absent classes require cautious interpretation.

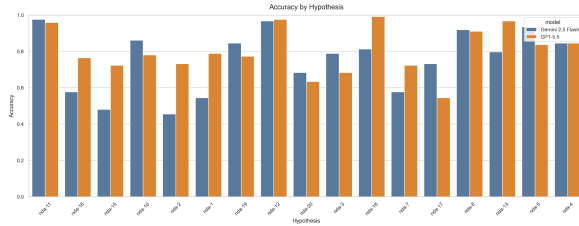


Figure 9: Accuracy by hypothesis.

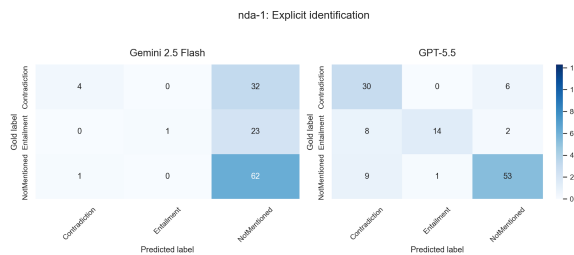


Figure 10: Per-hypothesis confusion counts. E = Entailment; C = Contradiction; NM = NotMentioned.

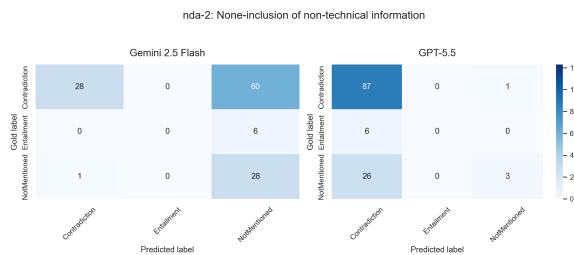


Figure 11: Per-hypothesis confusion counts. E = Entailment; C = Contradiction; NM = NotMentioned. (continued, panel 2)

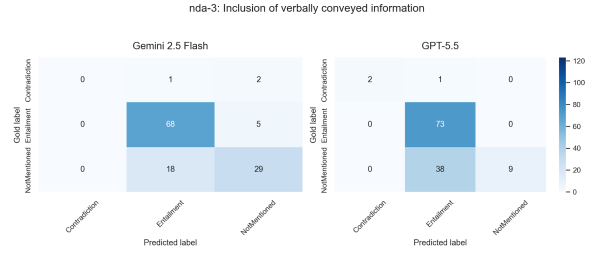


Figure 12: Per-hypothesis confusion counts. E = Entailment; C = Contradiction; NM = NotMentioned. (continued, panel 3)

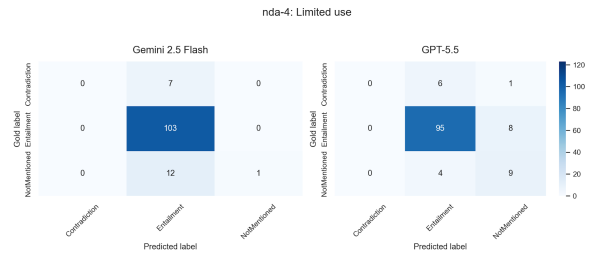


Figure 13: Per-hypothesis confusion counts. E = Entailment; C = Contradiction; NM = NotMentioned. (continued, panel 4)

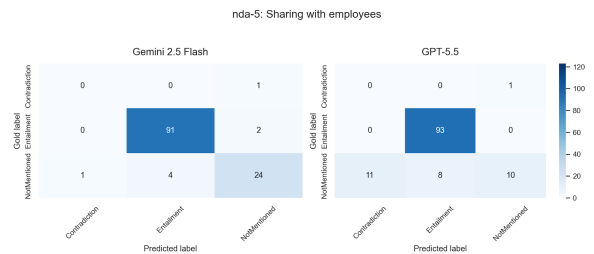


Figure 14: Per-hypothesis confusion counts. E = Entailment; C = Contradiction; NM = NotMentioned. (continued, panel 5)

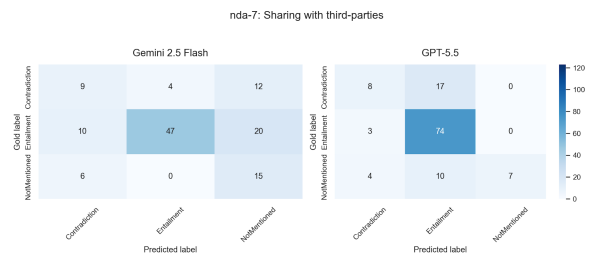


Figure 15: Per-hypothesis confusion counts. E = Entailment; C = Contradiction; NM = NotMentioned. (continued, panel 6)

ID	Legal proposition	Majority baseline	GPT-5.5	Gemini 2.5 Flash-Lite
nda-1	Explicit identification	51.2%	78.9% [†]	54.5%
nda-2	Technical-information limit	71.5%	73.2% [†]	45.5%
nda-3	Oral information	59.3%	68.3%	78.9%
nda-4	Limited use	83.7%	84.6%	84.6%
nda-5	Employee sharing	75.6%	83.7%	93.5% [†]
nda-7	Third-party sharing	62.6%	72.4%	57.7%
nda-8	Notice on compelled disclosure	62.6%	91.1%	91.9%
nda-10	Confidentiality of agreement	62.6%	78.0%	86.2%
nda-11	No reverse engineering	89.4%	95.9%	97.6%
nda-12	Independent development	59.3%	97.6%	96.7%
nda-13	Third-party acquirement	78.0%	96.7% [†]	79.7%
nda-15	No licensing	69.9%	72.4% [†]	48.0%
nda-16	Return/destruction on termination	69.1%	76.4%	57.7%
nda-17	Copying permission	60.2%	54.5%	73.2% [†]
nda-18	No solicitation	77.2%	99.2% [†]	81.3%
nda-19	Survival of obligations	64.2%	77.2%	84.6%
nda-20	Post-agreement retention	48.8%	63.4%	68.3%

Table 3: Complete hypothesis-level accuracy comparison and per-hypothesis majority baselines.

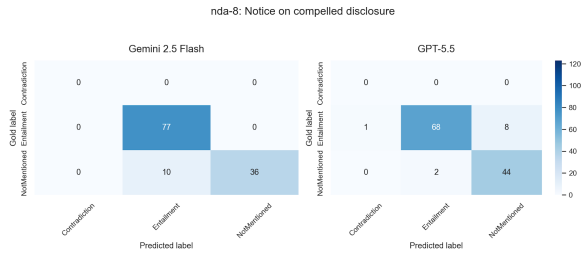


Figure 16: Per-hypothesis confusion counts. E = Entailment; C = Contradiction; NM = NotMentioned. (continued, panel 7)

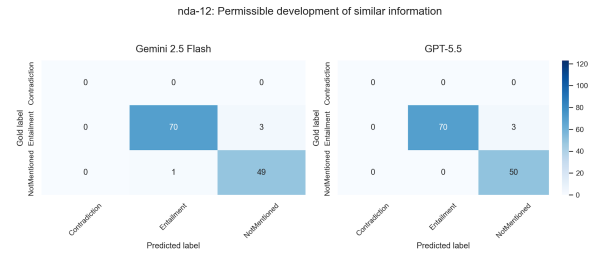


Figure 19: Per-hypothesis confusion counts. E = Entailment; C = Contradiction; NM = NotMentioned. (continued, panel 10)

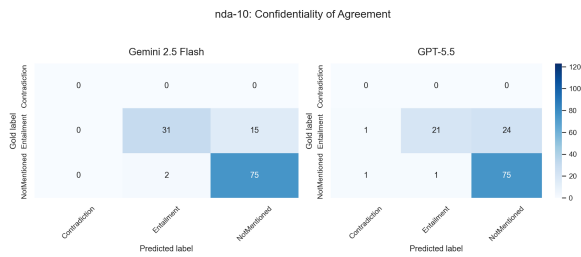


Figure 17: Per-hypothesis confusion counts. E = Entailment; C = Contradiction; NM = NotMentioned. (continued, panel 8)

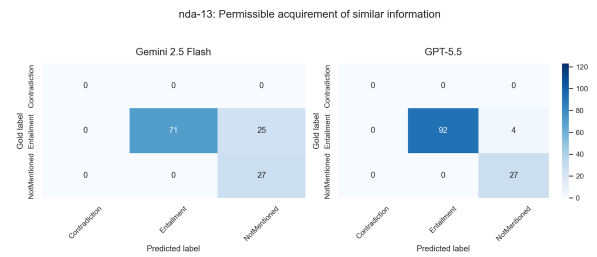


Figure 20: Per-hypothesis confusion counts. E = Entailment; C = Contradiction; NM = NotMentioned. (continued, panel 11)

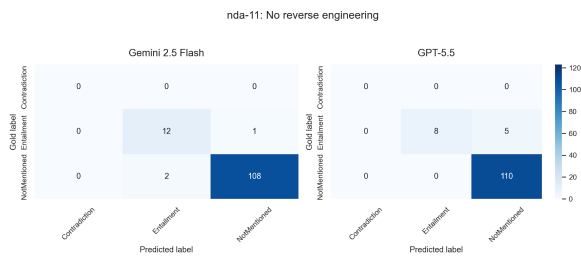


Figure 18: Per-hypothesis confusion counts. E = Entailment; C = Contradiction; NM = NotMentioned. (continued, panel 9)

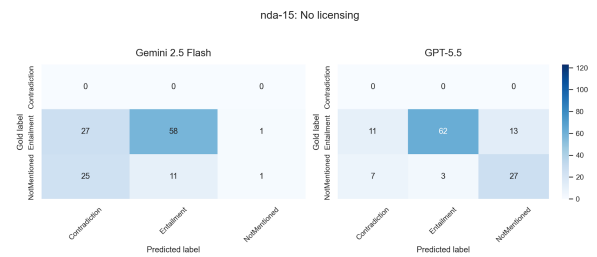


Figure 21: Per-hypothesis confusion counts. E = Entailment; C = Contradiction; NM = NotMentioned. (continued, panel 12)

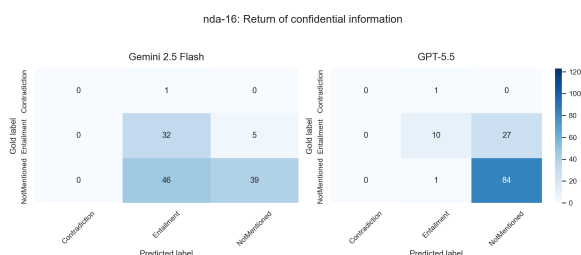


Figure 22: Per-hypothesis confusion counts. E = Entailment; C = Contradiction; NM = NotMentioned. (continued, panel 13)

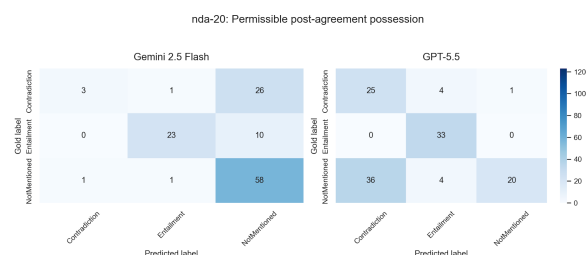


Figure 26: Per-hypothesis confusion counts. E = Entailment; C = Contradiction; NM = NotMentioned. (continued, panel 17)

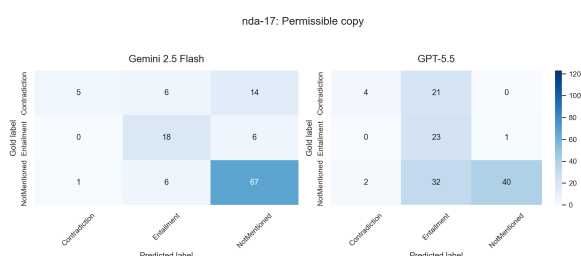


Figure 23: Per-hypothesis confusion counts. E = Entailment; C = Contradiction; NM = NotMentioned. (continued, panel 14)

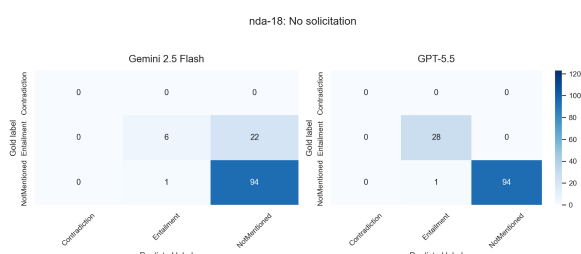


Figure 24: Per-hypothesis confusion counts. E = Entailment; C = Contradiction; NM = NotMentioned. (continued, panel 15)

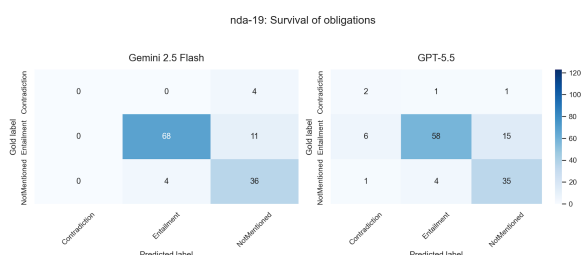


Figure 25: Per-hypothesis confusion counts. E = Entailment; C = Contradiction; NM = NotMentioned. (continued, panel 16)

A.1 Detailed Clause-Group Discussion

Performance varied substantially across the 17 hypotheses. The discussion below groups those results by the legal feature each hypothesis tests.

The first group concerns universal and exclusive scope: explicit identification (nda-1), technical-information limits (nda-2), and limitations on the use of confidential information (nda-4). Each turns on small but legally consequential terms such as all, only, and other than.

On nda-1, GPT achieved higher overall accuracy and much stronger Entailment and Contradiction recall, while Gemini retained higher NotMentioned recall. Gemini selected NotMentioned in 117 of 123 cases, showing limited responsiveness to the two minority labels.

On nda-2, neither system identified any of the six Entailments. GPT predicted Contradiction in 119 cases, correctly capturing 87 of 88 Contradictions but only 3 of 29 NotMentioned cases. Gemini displayed the reverse tendency: it captured 28 of 29 NotMentioned cases but only 28 of 88 Contradictions. The accuracy difference therefore reflects different default labels, not one model consistently understanding the hypothesis better.

On nda-4, both systems achieved 84.6%, yet neither identified any of the seven Contradictions. Gemini predicted Entailment in 122 of 123 cases; GPT was slightly more successful at recognising NotMentioned.

Taken together, these hypotheses show how high accuracy can arise from specialising in the dominant label rather than reliably distinguishing all three outcomes. GPT responded more often to substantive Entailment and Contradiction cases, while Gemini more often protected its score by selecting NotMentioned or the local majority label. Neither strategy handled universal and exclusive scope consistently.

The second group contains nda-3, which asks

whether confidential information may include information conveyed orally.

GPT identified all 73 Entailments but only 9 of 47 NotMentioned cases, achieving high Entailment recall by frequently assuming that oral information was included. Gemini was more balanced, correctly classifying 68 Entailments and 29 NotMentioned cases.

A recurring error arose when a contract required information to be orally identified or designated as confidential. That does not necessarily establish that the information itself may be conveyed orally. The systems made the same incorrect prediction in 17 cases, 16 of which were gold NotMentioned instances classified as Entailment.

This pattern is consistent with a lexical-trigger account: the word 'oral' may have been detected without preserving the distinction between oral designation and oral conveyance. Since the experiment does not observe internal reasoning, the explanation remains behavioural rather than mechanistic.

The third group concerns disclosure to employees (nda-5) and to third parties such as consultants, agents, and advisers (nda-7).

These hypotheses reproduce the broader response tendencies. Gemini was stronger on nda-5, correctly classifying 24 of 29 NotMentioned cases, while GPT identified every Entailment but often inferred employee-sharing permission where the contract said nothing. The pattern reversed for third-party disclosure: Gemini omitted many relevant provisions as NotMentioned, while GPT identified 74 of 77 Entailments. GPT, however, also classified 17 of the 25 Contradictions as Entailment.

The challenge was not merely to locate references to employees or advisers. The systems also had to determine whether disclosure depended on consent, a need-to-know requirement, equivalent confidentiality duties, or a narrow definition of authorised representatives, and whether those conditions were sufficient to satisfy the hypothesis. GPT tended to expand a limited exception into a broader permission; Gemini tended not to recognise the qualified permission at all.

The fourth group contains clauses with comparatively distinctive wording: notice before compelled disclosure (nda-8), confidentiality of the agreement itself (nda-10), reverse engineering (nda-11), independent development (nda-12), and independent third-party acquisition (nda-13).

Both systems generally performed well on this group, in which clauses often include recognisable

phrases such as 'court order', 'legal process', 'independently developed', or 'without use of confidential information'. Their strongest shared result was nda-12: GPT achieved 97.6% and Gemini 96.7%, against a majority baseline of 59.3%.

The pattern is consistent with distinctive wording making the relevant provision easier to locate, although the experiment cannot establish that either system relied on lexical matching alone or that standardised language caused the improvement. nda-13 was the principal exception: GPT was highly accurate, while Gemini missed several Entailments.

This group therefore suggests that both systems are most reliable when the relevant clause is stated directly and requires comparatively little inference.

The fifth group concerns negative polarity and contains nda-15, which asks whether the agreement grants no right or licence to the receiving party. The test set includes 86 Entailments and 37 NotMentioned cases, but no Contradictions.

GPT recognised more of the genuine no-licensing clauses, but still predicted Contradiction 18 times. Gemini was less balanced, selecting Contradiction 52 times and NotMentioned only twice despite 37 contracts that did not address licensing. The difficulty was therefore not simply the presence of negative language. The systems had to distinguish an explicit statement that no rights were granted from contractual silence on licensing.

Gemini often reversed the meaning of genuine no-licensing clauses while also missing many cases of silence. GPT handled the distinction more successfully, although a score close to the majority baseline still falls short of reliable discrimination across both labels.

More broadly, the negatively worded hypothesis altered the systems' entire label distribution. Terms such as nothing, not, and no rights sometimes appeared to trigger Contradiction even when the clause supported the proposition, while neither system consistently preserved the difference between a negative rule and an absent one.

The sixth group contains only nda-16, which concerns exact temporal triggers.

The seventh group contains nda-17, permission to copy. Gemini achieved 73.2% accuracy, compared with 54.5% for GPT. GPT's high Entailment rate appears to reflect an assumption that any reference to copies creates permission to make them. Gemini avoided many of these errors by preserving NotMentioned, but still missed a substantial share of Contradictions.

The eighth group contains nda-18, which tests whether an agreement prohibits the solicitation of employees or clients. No-solicitation clauses often use distinctive terms such as solicit, hire, employ, and recruit. GPT appears to have recognised these variants more consistently; Gemini was substantially less likely to classify the provision as present.

GPT was nearly perfect on nda-18. Gemini achieved 81.3% accuracy, only modestly above the 77.2% majority baseline, and detected 6 of the 28 Entailment cases.

The ninth group contains nda-19, survival of obligations after termination. Gemini achieved higher overall accuracy by performing better on the common Entailment and NotMentioned cases. It failed, however, to identify any of the four Contradictions; GPT identified two.

The final group contains nda-20, which asks whether the receiving party may retain some confidential information after the remainder has been returned or destroyed. Both systems struggled: Gemini achieved 68.3% accuracy and GPT 63.4%. Gemini was substantially stronger on NotMentioned cases, while GPT identified every Entailment and most Contradictions.

B Metric Definitions and Additional Diagnostics

This appendix preserves the statistical derivations and the figures removed from the main narrative. They remain important for replication and audit, although several repeat material presented more interpretively in the body of the paper.

B.1 Metric Definitions

After each prediction was compared with its gold label, overall accuracy and class-sensitive F1 metrics were calculated.

Raw aggregate accuracy was calculated as follows:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Number of Evaluated Instances}} \quad (1)$$

F1 scores were included because the label distribution was unequal and accuracy alone can conceal weak performance on the minority Contradiction class. Precision measures how often a prediction of a target label is correct; recall measures how many gold instances of that label are recovered. Both are calculated from true positives, false positives, and

false negatives. A true positive (TP) occurs when the predicted and gold labels match the target class. A false positive (FP) occurs when the target label is predicted but the gold label is one of the other two classes. A false negative (FN) occurs when the gold label is the target class but the model predicts another label.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$F_{1,\ell} = 2 \frac{\text{Precision}_\ell \text{Recall}_\ell}{\text{Precision}_\ell + \text{Recall}_\ell} \quad (4)$$

$$F_{1,\text{macro}} = \frac{F_{1,E} + F_{1,NM} + F_{1,C}}{3} \quad (5)$$

Balanced accuracy was calculated as the mean recall across the three labels:

Balanced accuracy = (Recall Entailment + Recall Contradiction + Recall NotMentioned) / 3

Overall macro F1 is well defined because all three labels appear in the complete test set. At the hypothesis level, however, several propositions contain no Contradiction cases. A fixed three-class macro F1 would assign the absent class a score of zero, capping the result at approximately 0.667 even when every observed instance was classified correctly. Hypothesis-level macro F1 is therefore interpreted alongside the confusion matrix and balanced accuracy over the labels actually present; it is not used alone to rank the 17 hypotheses by difficulty.

Accuracy and F1 can also conceal system-level behavioural failures when read only as aggregate statistics (Ribeiro et al., 2020). The evaluation therefore considers five levels of performance: overall, label-level, hypothesis-level, contract-level, and individual error patterns.

Confusion matrices were calculated for the complete test set and separately for each of the 17 hypotheses. They show both the magnitude and the direction of classification error.

The errors were also organised into three legally meaningful categories:

- Hallucinated provision: the gold label is NotMentioned, but the system predicts Entailment or Contradiction.
- Omitted provision: the gold label is Entailment or Contradiction, but the system predicts NotMentioned.

- Polarity reversal: the system predicts Entailment where the gold label is Contradiction, or Contradiction where the gold label is Entailment.

The taxonomy reflects the practical difference between these errors. Inventing a contractual permission is not equivalent to overlooking an obligation, even though each counts as one incorrect classification under accuracy.

B.2 Aggregate Distributions and Label-Level Metrics

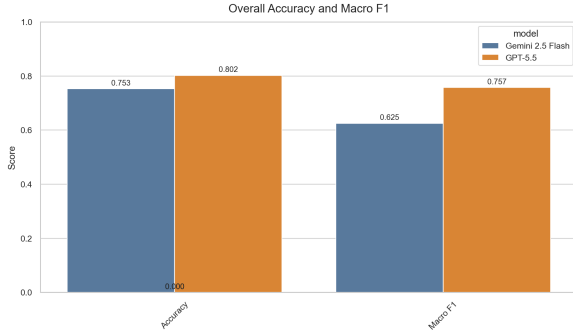


Figure 27: Original overall accuracy and macro F1 chart.

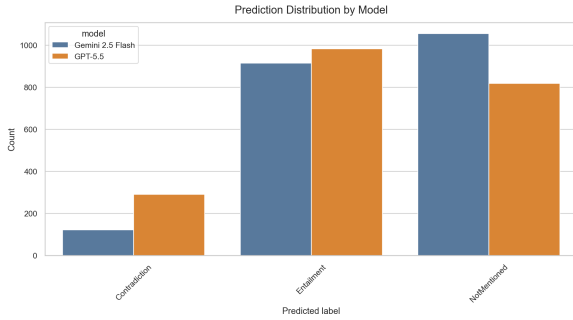


Figure 28: Original gold-label and model-prediction distribution chart.

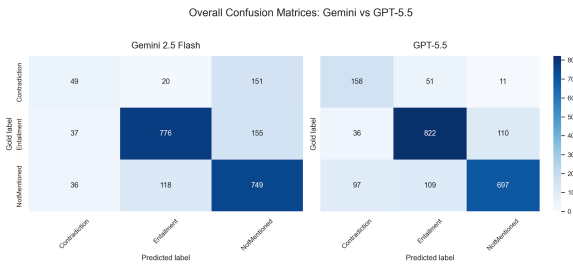


Figure 29: Overall confusion matrices in raw counts.

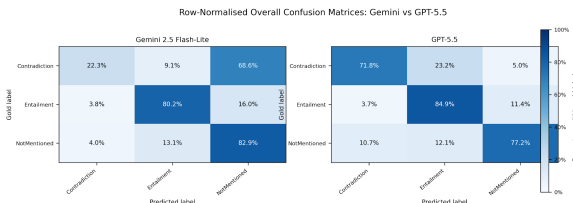


Figure 30: Row-normalised overall confusion matrices.

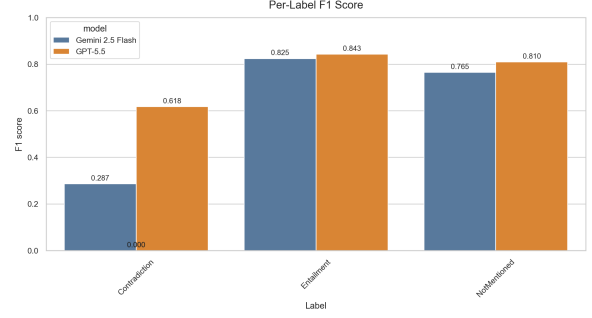


Figure 31: Label-level F1 scores.

B.3 Contract- and Hypothesis-Level Difficulty Statistics

One possible explanation for similar pockets of low accuracy is that some NDAs are inherently difficult. Were that the case, a contract that proved difficult for Gemini should also tend to prove difficult for GPT.

Each contract-level point was therefore defined as the proportion of the 17 hypotheses that one model classified correctly for a given NDA:

$$A_{m,i} = \frac{1}{17} \sum_{h=1}^{17} \mathbb{I}(\hat{y}_{m,i,h} = y_{i,h}) \quad (6)$$

where $\mathbb{I}(\cdot)$ equals 1 for a correct prediction and 0 otherwise. This yields 123 paired contract-level accuracies. At the hypothesis level, each point is the proportion of the 123 NDAs classified correctly for one legal proposition:

$$H_{m,h} = \frac{1}{123} \sum_{i=1}^{123} \mathbb{I}(\hat{y}_{m,i,h} = y_{i,h}) \quad (7)$$

For example, GPT-5.5 classified 97 of 123 documents correctly on nda-1 ($97/123 = 0.789$), while Gemini classified 67 correctly ($67/123 = 0.545$).

Pearson's coefficient was calculated from the paired accuracies as follows:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad (8)$$

Spearman's coefficient was calculated as the Pearson correlation of the average ranks:

$$\rho = \text{corr}(\text{rank}(x_i), \text{rank}(y_i)) \quad (9)$$

Here, x denotes Gemini accuracy and y denotes GPT-5.5 accuracy. Repeated accuracy values created ties, so average ranks were used rather than the common no-ties shortcut based on the sum of squared rank differences.

Gold label	System	Precision	Recall	F1
Entailment	GPT-5.5	0.837	0.849	0.843
Entailment	Gemini 2.5 Flash-Lite	0.849	0.802	0.825
Contradiction	GPT-5.5	0.543	0.718	0.618
Contradiction	Gemini 2.5 Flash-Lite	0.402	0.223	0.287
NotMentioned	GPT-5.5	0.852	0.772	0.810
NotMentioned	Gemini 2.5 Flash-Lite	0.710	0.829	0.765

Table 4: Label-level precision, recall, and F1 scores.

At the contract level ($n = 123$), Pearson $r = 0.074$ ($p = 0.419$) and Spearman $\rho = 0.090$ ($p = 0.322$). Both coefficients were close to zero and statistically non-significant, providing no meaningful evidence that the two systems consistently found the same NDAs difficult.

At the hypothesis level ($n = 17$), Pearson $r = 0.552$ ($p = 0.0216$) and Spearman $\rho = 0.628$ ($p = 0.00695$). These positive and statistically significant relationships indicated greater agreement about which legal propositions were relatively easy or difficult.

A binary correctness matrix supports the same conclusion by separating variation associated with contract identity, hypothesis identity, and their interaction.

In this descriptive decomposition, the largest component is the combined contract-hypothesis interaction and residual variation. Difficulty therefore arises from the proposition being tested, the way it is expressed or qualified in a particular NDA, and the model’s own response threshold and interpretive tendency.

There is nevertheless meaningful variation at the hypothesis level, supporting the narrower conclusion that some legal propositions were consistently more difficult for both systems.

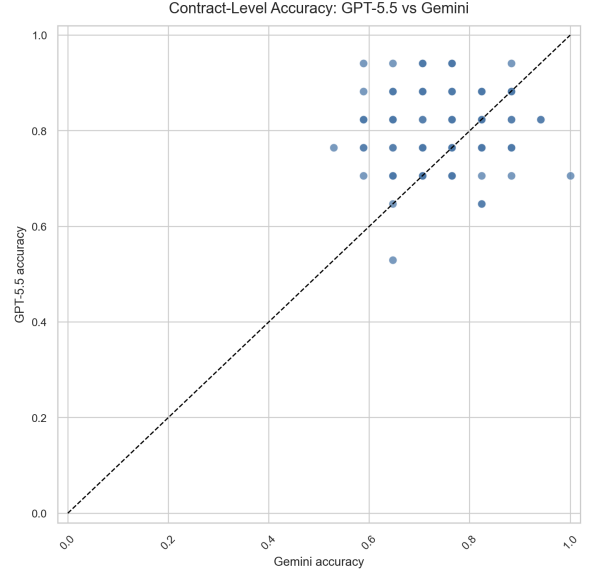


Figure 32: Contract-level accuracy comparison. The diagonal denotes equal performance.

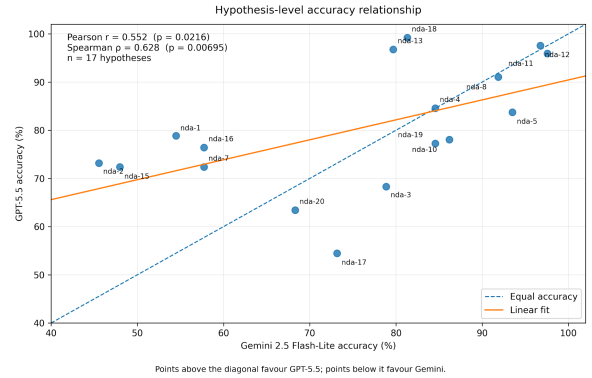


Figure 33: Hypothesis-level accuracy comparison, repeated from Figure 6. The dashed diagonal denotes equal performance; the solid line shows the least-squares linear fit.

B.4 Contract-Length Statistics

Contract length produced one of the clearest model-specific patterns. Gemini accuracy was negatively associated with both word count and character count.

Word count: Spearman $\rho = -0.378$, $p = 1.64 \times 10^{-5}$

Character count: Spearman $\rho = -0.389$, $p = 8.83 \times 10^{-6}$

No corresponding relationship was statistically significant for GPT.

To make the pattern more legible, the NDAs were also divided into quartiles by word count.

B.5 Model Agreement Diagnostics

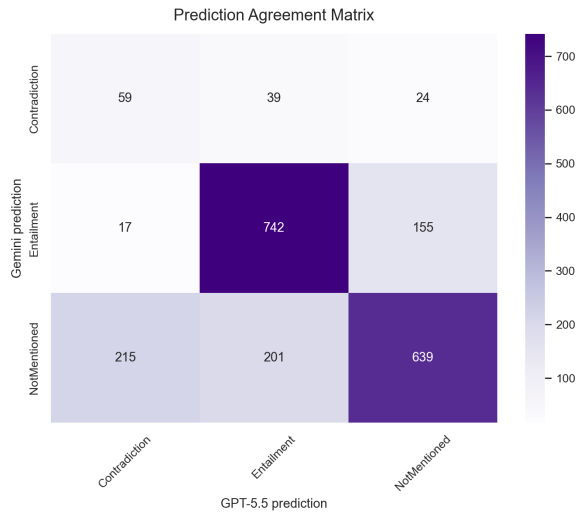


Figure 34: Agreement matrix between the two systems' predicted labels.

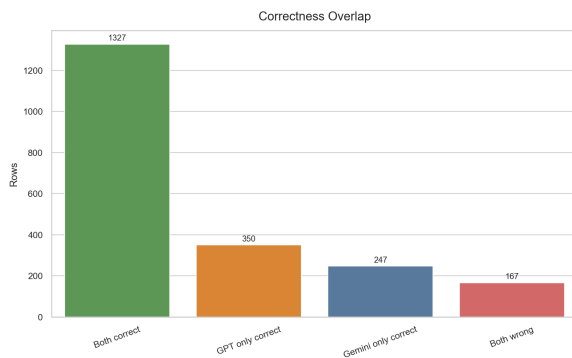


Figure 35: Correctness overlap between GPT-5.5 and Gemini 2.5 Flash-Lite.

Contract measure	Gemini ρ	Gemini p	GPT ρ	GPT p
Word count	-0.378	1.64×10^{-5}	0.129	0.154
Character count	-0.389	8.83×10^{-6}	0.126	0.166
Number of spans	-0.325	0.000246	0.034	0.706

Table 5: Correlation between contract-level difficulty measures and model accuracy.

Length group	Word-count range	GPT-5.5	Gemini 2.5 Flash-Lite	Difference
Shortest 25%	217-930	78.6%	79.7%	-1.1 pp
Second quartile	940-1,481	78.9%	75.5%	+3.4 pp
Third quartile	1,488-2,197	82.9%	76.1%	+6.9 pp
Longest 25%	2,229-6,348	80.5%	69.8%	+10.6 pp

Table 6: Accuracy by contract-length quartile.

C Additional Condition-Specific Results

C.1 Exception-Associated Hypotheses

Negation by exception, already identified as a challenge in ContractNLI, was associated with a substantial shift in performance (Koreeda and Manning, 2021). Across seven exception-associated hypotheses, both aggregate accuracy and macro F1 were lower than the corresponding means across all 17 hypotheses.

GPT’s aggregate accuracy advantage disappeared on the exception-associated subset: its accuracy fell by approximately 11.4 percentage points, compared with a decline of approximately 2.6 percentage points for Gemini. GPT retained the higher macro F1, so the result does not establish that Gemini handles exceptions better in general. It does show that both systems became less reliable and that a benchmark containing more exception-heavy items would yield a materially different headline comparison.

Negation by exception therefore makes the broader evaluation problem concrete. A model ranking is partly a property of the questions selected for the benchmark, and an aggregate advantage can disappear when greater weight is placed on one legal construction.

D Mechanism Hypotheses and Future Experiments

The results identify where the systems failed and the hypotheses under which those failures appeared. They cannot reveal the exact internal reasoning that produced them. Even so, the output patterns point towards four mechanisms that can be tested more directly in future experiments.

First, quantifier and scope reasoning. Words such as any, all, some, and only determine the reach of a provision. The universal- and exclusive-scope hypotheses show that a model can locate the relevant clause while still applying it too broadly or too narrowly.

Second, permission within a prohibition. Legal permissions are often written as narrow exceptions to a general restriction. Correct classification requires the model to preserve both levels of meaning, rather than expanding the exception into a general permission or overlooking it altogether.

Third, negative construction. The models sometimes appeared to associate words such as not, nothing, and no rights with the Contradiction label even when the clause entailed a negatively worded hypothesis. Controlled tests should therefore separate grammatical negation, logical contradiction, and contractual silence rather than treating them as one difficulty.

Fourth, temporal and conditional triggers. An obligation may apply only after termination, a written request, legal compulsion, or consent. The action itself can be easy to locate while the condition that gives it legal force is lost.

Taken together, these patterns suggest that the objective is not simply to add more reasoning to every instance. A useful model, or a useful routing system, would infer further when the document contains qualified or contradictory evidence while preserving NotMentioned when the relevant provision is absent. The 92.0% perfect-selector result gives a descriptive upper bound on that opportunity. A realistic system would still have to learn the routing decision without gold labels and be evaluated against an explicit model of legal error costs.

The most direct follow-up would compare

Subset	GPT accuracy	Gemini accuracy	GPT macro F1	Gemini macro F1
Exception-associated hypotheses	73.52%	73.75%	0.672	0.607
Remaining hypotheses	84.88%	76.34%	0.825	0.640

Table 7: Performance on exception-associated hypotheses and the remaining hypotheses.

matched variants of the same model family with reasoning enabled and disabled. Further experiments could require evidence spans, test calibrated abstention, and ask whether a learned router can preserve the complementary strengths of the two systems without reproducing their systematic errors.