

Do Long Contexts Help Legal Knowledge? A Case Study on US–China Securities Regulation

Stanford CS224N {Custom} Project

Yufei Peng
Stanford Law School
Stanford University
yfpeng@stanford.edu

Abstract

The ability of Large Language Models (LLMs) to answer legal question and handle long context have improved tremendously in the past years. Yet little do we know how updated LLMs are on new laws, how Retrieval Augmented Generation (RAG) would perform compared with models' parametric knowledge on legal task, and if different models would have a "home jurisdiction" bias that different LLMs developed and used in different jurisdictions have better performance on it's own domestic law. This study leverages the rapid growing legal regime surrounding the US - China capital market regulatory competition. Starting from 2019, both the US and China have enacted a series new laws and regulation targeting US-listed Chinese companies. By gathering the annual reports of these companies, which contains detailed legal risks and closely tracks legislative development of both countries, this research evaluate frontier LLMs' response in 6 different settings across English and Mandarin. The research find that all models perform substantially better under English prompting than Chinese, while clean RAG context yields heterogeneous gains and noisy full-document context disproportionately degrades fine-grained attribute accuracy despite comparable recall. We further observe an asymmetric home-jurisdiction effect that US-origin models show a recall advantage while Chinese-origin models do not. In the retrieval setting, although grounding is near-universal, 25% of verbatim-grounded responses still mis-extract drafting dates, revealing a gap between retrieval fidelity and structured attribute comprehension.

1 Key Information to include

- Mentor: Mirac Suzgun
- External Collaborators (if you have any): No
- Sharing project: No
- Late Days (if you are using any):

Team member	Late days remaining
<Yufei Peng>	<1>
Total pooled	<1>
Team size	<1>
Late days applied for the report	<1>

2 Introduction

Using AI to do legal research is a desirable goal for both the legal professions and computer engineers. With news like GPT-4 passed the bar exam [8], increasing researches have been done on benchmarking and evaluating models' legal reasoning capability[5, 14]. The bar exam for legal professions test on

comprehensive understanding of the foundational legal knowledge, and legal reasoning test on LLMs’ capacity to perform sophisticated reasoning tasks in domains like math or programming.

Recently, LLMs have been increasingly explored for regulatory intelligence tasks, including summarizing new rules, identifying applicable obligations, and answering legal queries. Prior work also recognizes that determining the operative legal requirements within a given domain is inherently challenging. In common-law systems, where judicial precedent can be binding, practitioners must sift through large bodies of case law to extract the relevant black-letter laws. Even in administrative law, which largely depends on agency rule-making, identifying “the law” remains difficult: agencies operate at multiple levels of government (e.g., federal and state), and their overlapping authority and evolving guidance complicate the construction of a coherent hierarchical regulatory picture[1, 9].

A natural question is whether frontier LLMs actually possess reliable knowledge of the specific regulatory instruments in a given domain, and whether grounding them in source documents improves or undermines that knowledge. Yet systematic evaluation of LLM regulatory recall—particularly across languages and document-context conditions—remains scarce. Existing work has largely focused on legal question answering or case-law retrieval in monolingual English settings, leaving the bilingual, multi-jurisdictional setting of cross-listed companies unstudied (See section 3.2). This study leverages the rapid growing legal regime surrounding the US - China capital market regulatory competition. Starting from 2019, both the US and China have enacted a series new laws and regulation targeting US-listed Chinese companies. For example, there are multiple US legislation, regulation and executive orders that target Chinese companies specifically. The most famous being the Holding Foreign Companies Accountable Act (HFCAA), which require US listed Chinese firms to delist if their accounting books are not given to Public Company Accounting Oversight Board(PCAOB) for inspection. Later on, there is a legislative attempt on passing Accelerating HFCAA, but that attempt failed in the Congress. On the Chinese side, there is new Anti-foreign Sanction Law, and the new ex-ante approval requirement before a Chinese company can go listing in a foreign financial market. These rapid changing legal regime can pose great confusion and challenges for LLMs. By gathering the annual reports of these companies, which contains detailed legal risks and closely tracks legislative development of both countries, this research evaluate frontier LLMs’ response in 6 different settings across two languages (English and Mandarin) and two very different and competing jurisdictions. This research aims to answer the following three questions: (1) can LLMs identify new laws in long, boilerplate-heavy disclosures, (2) to what extent does information quality impact the models’ ability to identify the documents, and (3) whether there is a jurisdictional "home country" advantage in LLMs.

3 Related Work

3.1 Information Retrieval

Information retrieval has been proved to help large pretrained language models in open-domain QA tasks[6]. With the development of LLMs, the open-domain question in parametric knowledge increased tremendously. Researchers further find that Retrieval Augmented Generation (RAG) techniques may help LLMs to identify new or more granular information that is not in the models parametric knowledge[4]. However, new research find that RAG can perform worse than zero-shot when giving misleading, conflicting, or only partially relevant retrieved content [12]. These findings motivate the 2 and 3 experiment settings. When LLMs being presented the same true data, but different in format. One being clean and more structured, the other being longer and noisier. For the current state-of-the-art LLMs, will data quality affect models’ retrieval ability, and if so, how much is that effect.

3.2 LLMs in Law

Evaluating LLMs ability in legal and finance tasks has been a great interests to many scholars. Researchers have tested LLMs’ ability to perform legal reasoning in both English and Chinese[5, 11, 2], and papers find both the potential and limitations of LLMs in legal applications, particularly in terms of challenges related to the interpretation of legal language and the accuracy of legal reasoning[10]. Other than legal interpretation, finding the law itself is a meaningful standalone task. Researchers find that multi-jurisdictional statutory analysis are difficult within the US [14]. Researchers in applying LLMs to legal domain also experimented using RAG to automate the tasks

that lawyers, policymakers, and researchers value in understanding legal requirements in a certain area of law[1, 9]. The same difficulty in finding legal document also exist in Chinese law, that most contemporary LLMs significantly struggle to faithfully reproduce legal provisions, as accurate retrieval and citation of legal provisions form the basis of legal reasoning, this deficiency critically undermines the reliability of models’ responses [11]. These research findings motivates me to set the task being legal provisions retrieval and focus on accuracy and faithfulness instead of legal reasoning.

Existing work in evaluating LLMs performance has largely focused on legal question answering or case-law retrieval in monolingual settings, leaving the bilingual, multi-jurisdictional setting of cross-listed companies unstudied. Existing research on comparative understanding between common law and civil law jurisdictions are scarce [13], comparing financial QAs across China and Dubai[13]. There is also research focus on multi-lingual task of German, French, Italian, Romansh, English within the Swiss jurisdiction[7]. This study fills this gap by evaluate frontier LLMs’ response in 6 different settings across two languages (English and Mandarin) and two very different and competing jurisdictions. Adding on the home jurisdiction test of models that testing whether models from different origin would perform better in their home jurisdiction assuming more familiarity in training data in their own domestic jurisdiction.

4 Approach

This research study models’ capability outputting legal provisions for a rapidly evolving US–China capital market legal regime. I constructed a pipeline to do the evaluation. From US-China Economic and Security Review Commission, I gathered the list of current list on US listed Chinese companies, looked up their CIK number (a unique 10-digit identifier assigned by the US Securities and Exchange Commission (SEC) to companies, and then leveraged SEC’s EDGAR database to collect all of the annual reports of the current listing Chinese companies from the list. Public company’s annual reports are lengthy and technical. Batch-downloaded data are noisy with nonuniform html annotations. Risk factor section contains all the legal documents. For the retrieval-clean condition, I manually-stripped documents into the risk factor section and then prepend a single document to the user prompt. For the retrieval-noisy condition, I concatenated 2019-2021, 3 years of full 20-F documents (stripped financial table in order to fit the model context window) with the html annotations into a single document to the user prompt.

Given each natural-language query prompt, a LLM produces (i) a list of relevant legal instruments (laws, regulations, executive orders, measures, etc.) and (ii) structured attributes for each instrument, including jurisdiction (US or China), a regulator (responsible authority), and key dates (drafting date and/or effective date, when applicable). Under retrieval conditions, the model is additionally asked to produce a supporting evidence quote drawn from a provided long context. For querying the model, I coded a pipeline to using the prompts to query models, supplementing manually collected data via the user interface for other models without API access. (See section 5.1) All model outputs are required to be valid JSON and are stored in a unified dataset.

For each model response, I extract a set of outputted instruments and fields. I constructed 2 dictionaries for standardizing the model outputs. (1) Instrument name. Model responses listing instrument names in various format. These names are mapped to a controlled canonical set via substring patterns, applied separately to English and Chinese name fields, then match grouped Chinese names to the canonical English names. (2) Regulators. Model responses of regulators in various format as well. I created a dictionary standardizing regulators to facilitate evaluation.

Ground Truth Construction. Given canonical instrument names and unmatched instrument names, I manually checked with Lexis, Federal Registry, PKULaw¹ and various other sources, such as SEC website for accurate date information to establish the ground truth for the evaluation.

Human Evaluation. I choose 2019-2021 to test on information retrieval because it was the period that most new laws surrounding the US-China financial markets are being drafted and proposed. For my first setting to test on parametric knowledge, the dates of legal instruments should match my ground truth to be considered accurate. But in retrieval tasks, if the model output date that is correct to my ground-truth but incorrect as to the document, that would be considered incorrect as parametric

¹<https://www.pkulaw.com/>

Table 1: Grading Rubric for Date Attributes (left) and Grounding Quotes (right).

Date Attribute Score			Grounding Quote Score		
Score	Category	Criterion	Score	Label	Criterion
2	Correct	Exact match to quote	1	Grounded	Quote locatable in source filing
1	incomplete	Not capturing specific date/month in quote	0	Ungrounded	Quote absent from source
0	neutral	left it blank as not mentioned in quote	<i>Layer 2 — Faithfulness (grounded = 1)</i>		
-1	Far wrong		3	Verbatim	Direct copy, all key terms preserved
			2	Partial	Accurate but omits material clauses
			1	Paraphrase	Correct meaning, substantially re-worded
			0	Distorted	Wording sourced but meaning altered

Table 2: 2×3 factorial prompt design. All six prompts share the same extraction schema; retrieval prompts additionally enforce a closed-world constraint and require `source_quote` anchoring.

ID	Language	Condition	Description
a	English	Parametric	No document provided. Model draws on parametric knowledge only. No closed-world constraint; no <code>source_quote</code> field.
b	English	Clean retrieval	Risk factors section only (pre-isolated, low noise). Closed-world constraint: instruments must be named in the document. Requires <code>source_year</code> and <code>source_quote</code> .
c	English	Noisy retrieval	Full 20-F filing. Same constraints as b; adds <code>source_section</code> field to capture location within the filing.
d	Chinese	Parametric	Chinese translation of a. Identical schema and scope; no document or grounding fields.
e	Chinese	Clean retrieval	Chinese translation of b. Risk factors section with closed-world constraint and <code>source_quote</code> requirement.
f	Chinese	Noisy retrieval	Chinese translation of c. Full 20-F with closed-world constraint, <code>source_section</code> , and <code>source_quote</code> .

knowledge leaked to the response in that setting. To evaluate retrieval quality, I manually evaluated retrieval quality with the grading rubric. 1

5 Experiments

The experiment tests on four factors:

1. **RQ1: Parametric legal knowledge.** What does the model know without external retrieval?
2. **RQ2: Effect of external information** Does providing additional context help or impede correctness on legal knowledge?
3. **RQ3: Effect of information format.** Does the format (clean vs noisy context) affect understanding and evidence use?
4. **RQ4: Effect of model origin.** Does the model origin (US-based vs China-based) affect legal provisions retrieval in different jurisdictions?

There are six prompt templates (a–f) 2 that operationalize these factors:

- **Parametric probing:** (a) English, no context; (d) Chinese, no context.
- **Retrieval probing with clean context:** (b) English + clean context; (e) Chinese + clean context.

- **Retrieval probing with noisy context:** (c) English + noisy context; (f) Chinese + noisy context.

Prompts (a,d) primarily test RQ1. Comparisons (b vs a) and (e vs d) test the benefit of external information(RQ2), while (c vs b) and (f vs e) quantify the penalty of noise (RQ3). Paired comparisons (d vs a), (e vs b), and (f vs c) isolate language effects (RQ4).

5.1 Data

My data started with 257 firms’ past annual reports (Form 20-F). These are the companies that identified as the universe of Chinese companies listed in the US, ² that also maintain an active CIK number with the Securities Exchange Commission as of February 2026. All of the annual reports adds up to 1500 files. After filtering to firms have consistent 2019 to 2021 reports, there are 85 firms left, with 255 files. During experiment, I did further sampling on the data due to resource constraint. For details of sampling, see Section 5.4.

5.2 Evaluation method

I evaluate model outputs at two levels. First, I evaluate whether the model retrieves the correct legal instruments. Let G denote the gold set of legal instruments for a condition, and let $\hat{P}$ denote the canonicalized predicted set. I compute set-based precision, recall, and F1 as

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{|G|}, \quad F1 = \frac{2PR}{P + R}. \quad (1)$$

Recall is the main metric because a model can obtain high precision by naming only a small number of provisions, while the task requires broad coverage of the relevant legal inventory. F1 is reported as a balanced summary, and precision is included for completeness. All three metrics are computed on canonicalized sets rather than raw lists, so repeated mentions do not increase performance scores.

These retrieval metrics are used to address the main experimental comparisons. For RQ1, I evaluate performance under the no-context prompts. For RQ2, I compare clean-context prompts against their no-context counterparts within the same model. For RQ3, I compare noisy-context prompts against clean-context prompts within the same model. For RQ4, I split the ground truth by legal jurisdiction and compute retrieval performance separately for the two subsets. This makes it possible to examine whether model-origin groups differ in their relative coverage of the US and Chinese law under the same prompt condition.

Secondly, because retrieval accuracy alone does not show whether a model is actually using the provided document, I also conduct a manual evaluation of supporting evidence. This evaluation is applied to the quoted support and date attributes attached to predicted legal instruments. ¹ The purpose is to distinguish correct retrieval from grounded and faithful use of retrieved material.[3]

5.3 Experimental details

I evaluate five LLMs spanning US-based and China-based model families:

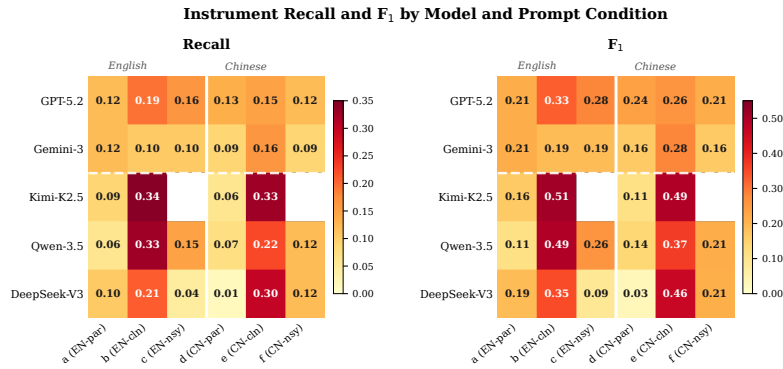
- **US-based:** gemini3-thinking, gpt5.2-thinking.
- **China-based:** kimi-k2.5, qwen3.5plus-thinking, deepseekv3.2.

All models are queried with the same prompt templates (a–f). After sampling from sector and VIE status (see appendix), I filtered 12 firms to strip to produce clean documents and queried Kimi on retrieval probing with clean document setting. Due to API access constraints, I queried all 12 firms through Kimi’s API, while manually collected all other responses from models using Alibaba’s annual reports in the retrieval settings I choose Alibaba as the input document for two reasons: (1) it has the biggest market cap among all Chinese issuers in 2025, and (2) in my robustness check with Kimi, Alibaba has the most retrieved legal instruments compared with all other companies(see appendix). Kimi-k2.5, although being the model able to handle the longest context in the model family, it’s still unable to process noisy document chunked each year due to token restriction. I avoid further chunk

²<https://www.uscc.gov/research/chinese-companies-listed-major-us-stock-exchanges>

Table 3: Main Result on Legal Instrument across Models and Prompts

Model	Condition	Count	TP	FP	Prec.	Rec.	F1
DeepSeek-V3	EN param	7	7	0	1.000	0.104	0.189
	EN clean	14	14	0	1.000	0.209	0.346
	EN noisy	3	3	0	1.000	0.045	0.086
	CN param	1	1	0	1.000	0.015	0.029
	CN clean	20	20	0	1.000	0.299	0.460
	CN noisy	9	8	1	0.889	0.119	0.211
Gemini-3	EN param	8	8	0	1.000	0.119	0.213
	EN clean	7	7	0	1.000	0.104	0.189
	EN noisy	7	7	0	1.000	0.104	0.189
	CN param	6	6	0	1.000	0.090	0.164
	CN clean	11	11	0	1.000	0.164	0.282
	CN noisy	6	6	0	1.000	0.090	0.164
GPT-5.2	EN param	8	8	0	1.000	0.119	0.213
	EN clean	13	13	0	1.000	0.194	0.325
	EN noisy	12	11	0	1.000	0.164	0.282
	CN param	9	9	0	1.000	0.134	0.237
	CN clean	11	10	0	1.000	0.149	0.260
	CN noisy	9	8	0	1.000	0.119	0.213
Kimi-K2.5	EN param	6	6	0	1.000	0.090	0.164
	EN clean	24	23	0	1.000	0.343	0.511
	CN param	4	4	0	1.000	0.060	0.113
	CN clean	31	22	0	1.000	0.328	0.494
Qwen-3.5	EN param	4	4	0	1.000	0.060	0.113
	EN clean	22	22	0	1.000	0.328	0.494
	EN noisy	10	10	0	1.000	0.149	0.260
	CN param	5	5	0	1.000	0.075	0.139
	CN clean	15	15	0	1.000	0.224	0.366
	CN noisy	8	8	0	1.000	0.119	0.213

Figure 1: Instrument Recall and F₁ by Model and Prompt Condition

noisy document into shorter context, since RQ3 is testing on models’ ability to process noisy long context, I left Kimi-k2.5 in prompt c and f blank.

5.4 Results

Table 3 and Figures 1 1 show four main results. First, precision is near ceiling for most model–prompt pairs, since the ground truth is the union of all outputted legal instrument, while recall remains modest. This means the main difficulty is not naming many hallucinated laws, but recovering enough of the relevant legal instruments. Under no-context prompts, recall is low across models, which suggests that parametric legal knowledge alone is limited and not as comprehensive.

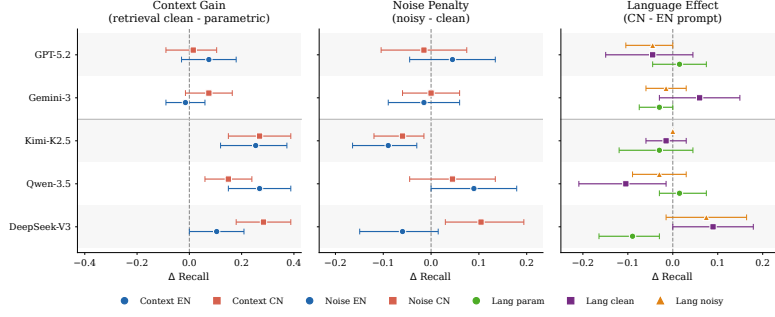


Figure 2: Paired Delta Recall

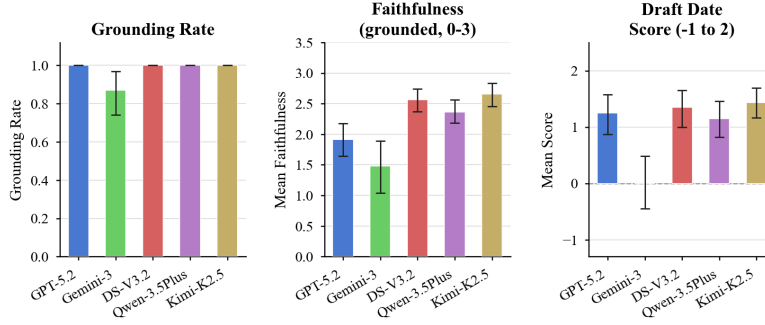


Figure 3: Evidence Quality by Model (95 bootstrap CI)

Second, clean retrieved context generally improves retrieval. In the paired-delta plots and the ΔF_1 heatmap, most models show positive changes from parametric to clean conditions, especially Kimi-K2.5, Qwen-3.5, and DeepSeek-V3.2. Third, noisy context usually reduces performance relative to clean context. This drop is visible for most models and is especially large for DeepSeek-V3 and Qwen-3.5, while GPT-5.2 declines less.

Fourth, retrieval patterns differ by legal jurisdiction. In the jurisdiction-gap plots, Chinese-origin models show a more positive gap under Chinese clean prompts, indicating relatively stronger recall on Chinese law than US law in that setting. US-origin models are closer to balanced or slightly tilted toward US law.

The human evidence evaluation adds a second layer. Grounding rates are high overall, often close to 1.0, which means that quoted support is usually locatable in the source. However, faithfulness and date quality vary more than grounding alone. By model, Kimi-K2.5 and DeepSeek-V3 have the strongest average faithfulness and date scores, while Gemini-3 is weaker and more variable. By condition, clean prompts outperform noisy prompts on faithfulness and draft-date scores in both languages, with the clearest drop under Chinese noisy prompts.

Taken together, the results suggest three broad findings. First, legal-instrument retrieval from parametric memory alone is limited. Second, clean retrieval context is generally helpful, but noisy context often weakens that benefit. Third, the effect of language and legal jurisdiction is not uniform across models, and appears to interact with model origin.

6 Analysis

These results suggest that retrieval success and evidence quality are related but not identical. A model can retrieve the correct instrument yet still provide weaker support in the quote or date fields. This is why retrieval F_1 alone is not enough. The human evaluation shows that most models can point to some source text, but the quality of that use still depends on the model and the prompting condition.

The contrast between clean and noisy settings is especially important. Clean context improves both retrieval coverage and evidence quality, while noisy context lowers not only recall and F_1 , but also quote faithfulness and date accuracy. These evidences suggest that models do not benefit from more text by itself, but only benefit from clean and structured text. The jurisdiction results point to a similar pattern: model behavior depends not only on overall capability, but also fluctuate on the alignment between model origin, prompt language, and legal jurisdiction.

Overall, Kimi are strongest on clean context retrieval, DeepSeek shows large gains when using Chinese prompts, and Qwen performs better in retrieval under English prompts. GPT-5.2 is the most stable across all conditions but less strong at peak performance, and Gemini-3 is the weakest and least consistent across settings.

There are also several limitations to these results. First, the number of models is small, so the model-origin comparisons should be read as descriptive rather than definitive. Second, Kimi does not have complete outputs for every prompt condition, which limits direct comparison in a few cells. Third, the current results focus mainly on legal-instrument retrieval itself. The human evaluation of quote grounding, faithfulness, and date support gives a fuller picture of output quality, but it is secondary to the main retrieval metrics and is best interpreted alongside selected examples.

7 Conclusion

This project examined legal-instrument retrieval under six prompt conditions designed to test parametric knowledge, the value of external information, the effect of noisy evidence, and differences associated with model origin. The main findings are threefold.

First, parametric legal knowledge alone is limited. Across models, no-context prompting retrieves only a small share of the relevant legal instruments. Second, external context is generally helpful when it is clean, but this benefit often weakens once the evidence becomes noisy. This suggests that retrieval quality in legal tasks depends not only on whether relevant material is present, but also on whether it is presented in a form that models can reliably use. Third, retrieval behavior differs across legal jurisdictions, and these differences appear to be related to model origin. In particular, Chinese-origin models show a clearer relative advantage on Chinese-law retrieval under Chinese clean-context prompts, while US-origin models are more balanced or relatively stronger on US law.

A main contribution of the project is the combination of set-based retrieval evaluation with manual assessment of grounding, faithfulness, and document-supported dates. This makes it possible to separate two questions that are often blurred together: whether a model retrieves the right legal instrument, and whether it uses the retrieved document accurately. That distinction is important for legal applications, where unsupported or weakly grounded evidence can be as problematic as retrieval failure itself.

At the same time, the study has clear limitations. The number of evaluated models is small, and the model-origin comparison should therefore be interpreted with caution. The legal domain is also narrow, focusing on a specific period and topic area rather than the full range of US and Chinese law. In addition, the current experiments evaluate retrieval quality at the instrument level, while broader legal reasoning over these materials remains outside the scope of the project, which calls for future work.

References

- [1] Mohamed Afane et al. “Benchmarking Legal RAG: The Promise and Limits of AI Statutory Surveys”. en. In: (2026).
- [2] Ilias Chalkidis et al. *LexGLUE: A Benchmark Dataset for Legal Language Understanding in English*. arXiv:2110.00976 [cs]. Nov. 2022. DOI: 10.48550/arXiv.2110.00976. URL: <http://arxiv.org/abs/2110.00976> (visited on 03/05/2026).
- [3] Shahul Es et al. “RAGAs: Automated Evaluation of Retrieval Augmented Generation”. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Ed. by Nikolaos Aletras and Orphee De Clercq. St. Julians, Malta: Association for Computational Linguistics, Mar. 2024, pp. 150–158. DOI: 10.18653/v1/2024.eacl-demo.16. URL: <https://aclanthology.org/2024.eacl-demo.16/> (visited on 03/14/2026).

- [4] Wenqi Fan et al. “A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models”. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD ’24. New York, NY, USA: Association for Computing Machinery, Aug. 2024, pp. 6491–6501. ISBN: 979-8-4007-0490-1. DOI: 10.1145/3637528.3671470. URL: <https://dl.acm.org/doi/10.1145/3637528.3671470> (visited on 03/04/2026).
- [5] Neel Guha et al. “LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models”. en. In: *Advances in Neural Information Processing Systems* 36 (Dec. 2023), pp. 44123–44279. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/89e44582fd28ddfea1ea4dcb0ebbf4b0-Abstract-Datasets_and_Benchmarks.html (visited on 03/02/2026).
- [6] Patrick Lewis et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. arXiv:2005.11401 [cs]. Apr. 2021. DOI: 10.48550/arXiv.2005.11401. URL: <http://arxiv.org/abs/2005.11401> (visited on 03/10/2026).
- [7] Joel Niklaus et al. “LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 3016–3054. DOI: 10.18653/v1/2023.findings-emnlp.200. URL: <https://aclanthology.org/2023.findings-emnlp.200/> (visited on 03/13/2026).
- [8] Stanford Law School. *GPT-4 Passes the Bar Exam: What That Means for Artificial Intelligence Tools in the Legal Profession*. en. Apr. 2023. URL: <https://law.stanford.edu/2023/04/19/gpt-4-passes-the-bar-exam-what-that-means-for-artificial-intelligence-tools-in-the-legal-industry/> (visited on 03/13/2026).
- [9] Faiz Surani et al. “What Is the Law? A System for Statutory Research (STARA) with Large Language Models”. In: *Proceedings of the Twentieth International Conference on Artificial Intelligence and Law*. ICAIL ’25. New York, NY, USA: Association for Computing Machinery, Jan. 2026, pp. 394–398. ISBN: 979-8-4007-1939-4. DOI: 10.1145/3769126.3769219. URL: <https://dl.acm.org/doi/10.1145/3769126.3769219> (visited on 03/02/2026).
- [10] Jiaqi Wang et al. *Legal Evaluations and Challenges of Large Language Models*. arXiv:2411.10137 [cs]. Nov. 2024. DOI: 10.48550/arXiv.2411.10137. URL: <http://arxiv.org/abs/2411.10137> (visited on 03/03/2026).
- [11] Xinzhe Xu et al. *CLaw: Benchmarking Chinese Legal Knowledge in Large Language Models - A Fine-grained Corpus and Reasoning Analysis*. arXiv:2509.21208 [cs]. Sept. 2025. DOI: 10.48550/arXiv.2509.21208. URL: <http://arxiv.org/abs/2509.21208> (visited on 03/05/2026).
- [12] Linda Zeng et al. *Worse than Zero-shot? A Fact-Checking Dataset for Evaluating the Robustness of RAG Against Misleading Retrievals*. arXiv:2502.16101 [cs]. Jan. 2026. DOI: 10.48550/arXiv.2502.16101. URL: <http://arxiv.org/abs/2502.16101> (visited on 03/02/2026).
- [13] Xinyan Zhang et al. “RIRAG+: A Cross-Jurisdictional Financial Legal QA Framework Bridging Civil and Common Law Systems”. In: *2025 25th International Conference on Software Quality, Reliability, and Security Companion (QRS-C)*. ISSN: 2693-9371. July 2025, pp. 11–20. DOI: 10.1109/QRS-C65679.2025.00013. URL: <https://ieeexplore.ieee.org/abstract/document/11216537> (visited on 03/03/2026).
- [14] Lucia Zheng et al. “A Reasoning-Focused Legal Retrieval Benchmark”. In: *Proceedings of the Symposium on Computer Science and Law on ZZZ*. arXiv:2505.03970 [cs]. Mar. 2025, pp. 169–193. DOI: 10.1145/3709025.3712219. URL: <http://arxiv.org/abs/2505.03970> (visited on 03/03/2026).

A Appendix (optional)

3

³There are 29 companies that are identified as Chinese issuers in 2025 but I can no longer match their Symbol(ticker) to an active CIK number anymore. It could be due to merger and acquisition or de-listing from the US stocks.

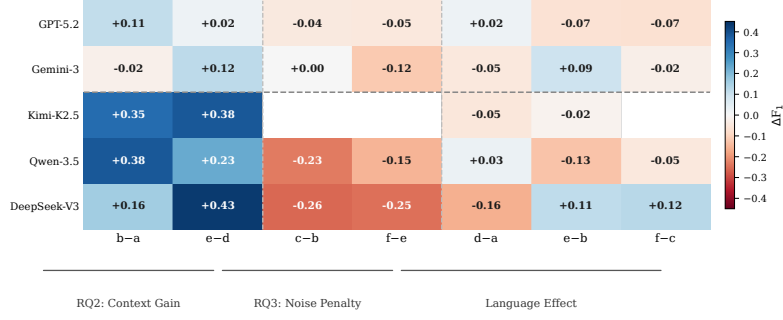


Figure 4: ΔF_1 Heatmap: Context Gain, Noise Penalty, and Language Effect

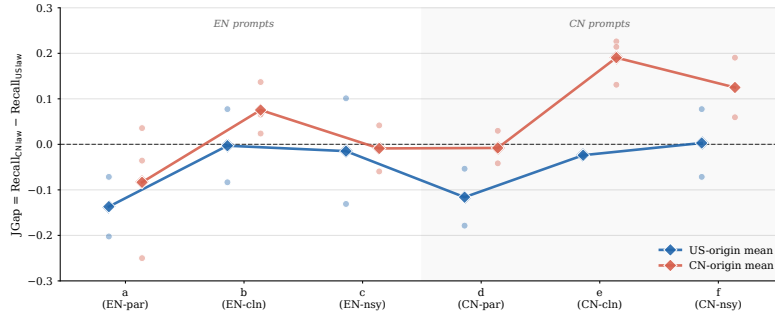


Figure 5: Jurisdiction Gap (JGap) by Prompt Condition and Model Origin

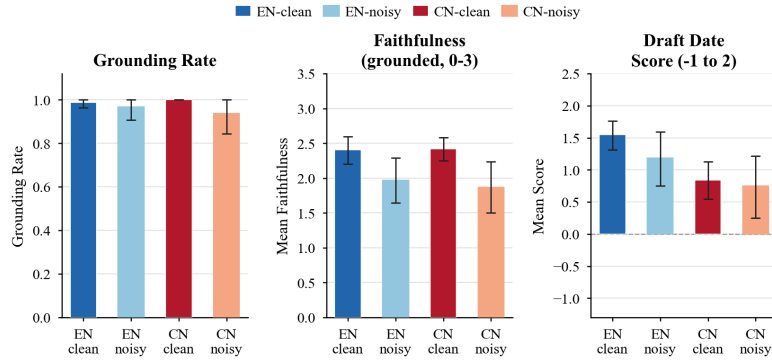


Figure 6: Evidence Quality by Condition (95 bootstrap CI)

Table 4: Sector and VIE structure of the 85-company eligible pool (US-listed Chinese companies with complete 2019–2021 20-F filings).

Sector	Non-VIE	VIE	Total
Technology	8	18	26
Consumer Services	4	16	20
Finance	1	11	12
Business Services	1	4	5
Capital Goods	3	2	5
Consumer Durables	2	2	4
Health Care	3	1	4
Basic Industries	3	0	3
Consumer Non-Durables	2	1	3
Basic Materials	1	0	1
Energy	0	1	1
Transportation	0	1	1
Total	28	57	85

Table 5: Final 12-company purposive sample, stratified by sector group (Technology / Consumer) and VIE structure. Companies selected by descending market capitalisation within each stratum.

Ticker	Company	Sector	VIE	IPO	Market Cap (\$M)
<i>Consumer Non-VIE</i>					
HTHT	H World Group Limited	Consumer Services	No	Mar 2010	11,900
MSC	Studio City International Holdings	Consumer Services	No	Oct 2018	738
DOGZ	Dogness (International) Corporation	Consumer Durables	No	Dec 2017	360
<i>Consumer VIE</i>					
JD	JD.com, Inc.	Consumer Services	Yes	May 2014	67,875
NIO	NIO Inc.	Consumer Durables	Yes	Sep 2018	9,670
VIPS	Vipshop Holdings Limited	Consumer Services	Yes	Mar 2012	8,821
<i>Technology Non-VIE</i>					
DQ	Daqo New Energy Corp.	Technology	No	Oct 2010	1,253
JKS	JinkoSolar Holding Co., Ltd.	Technology	No	May 2010	1,117
CSIQ	Canadian Solar Inc.	Technology	No	Nov 2006	678
<i>Technology VIE</i>					
BABA	Alibaba Group Holding Limited	Technology	Yes	Sep 2014	345,960
NTES	NetEase, Inc.	Technology	Yes	Jun 2000	68,655
BIDU	Baidu, Inc.	Technology	Yes	Aug 2005	33,399

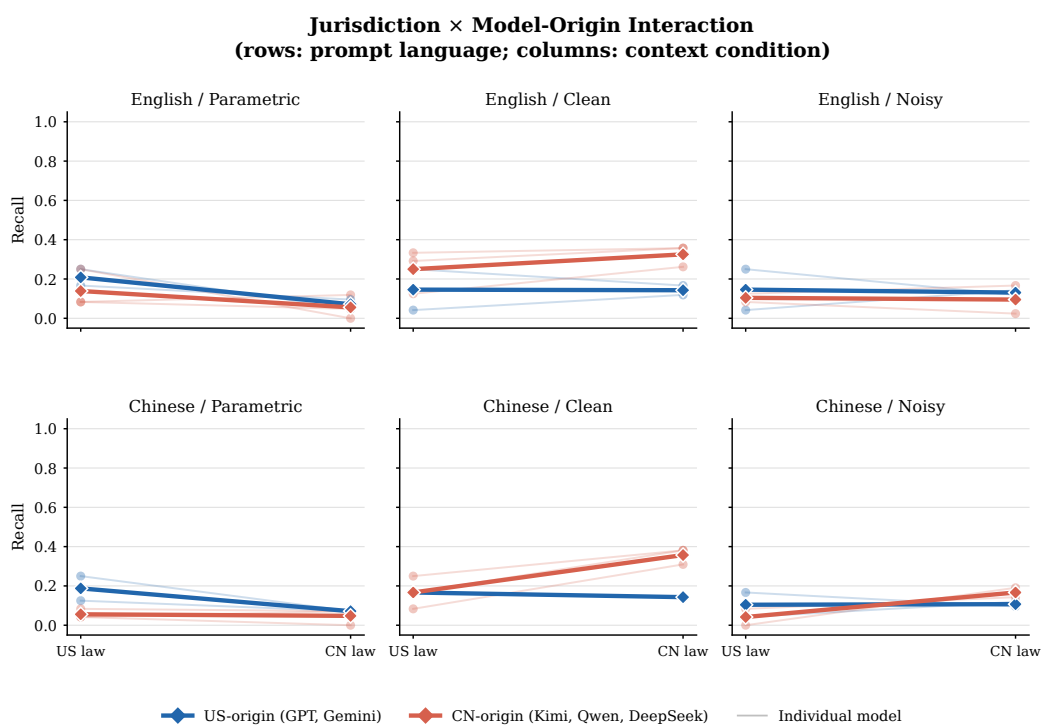


Figure 7: jurisdiction *Model-Origin Interaction Plots