GAVEL: AGENT MEETS CHECKLIST FOR EVALUATING LLMS ON LONG-CONTEXT LEGAL SUMMARIZATION

Anonymous authors
Paper under double-blind review

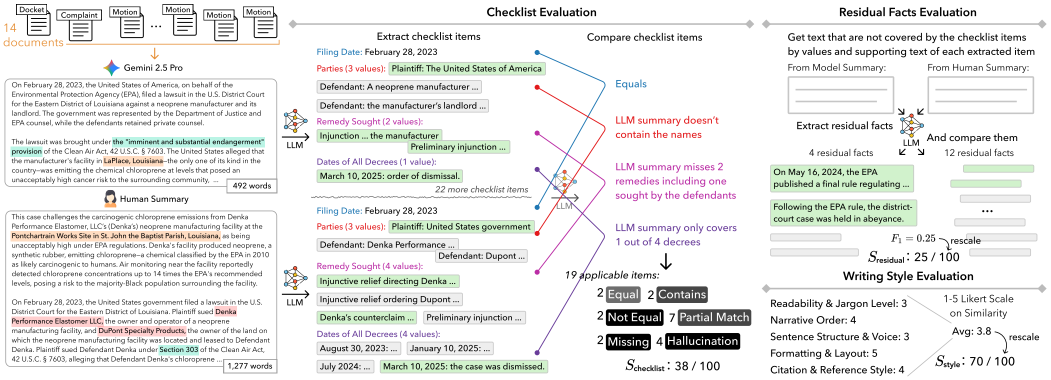
ABSTRACT

Large language models (LLMs) are increasingly applied in legal practice, with case summarization being a key long-context task where cases often exceed 100K tokens across multiple documents. Existing evaluation methods rely on checklist comparisons but use coarse-grained extraction that merges multiple values into single text blocks, missing partial matches when comparing them. They also overlook content beyond predefined checklist categories and lack writing style evaluation. In this paper, we introduce GAVEL-REF, a reference-based evaluation framework that improves checklist evaluation through multi-value extraction with supporting text, and further incorporates residual fact and writing-style assessments. Using GAVEL-REF, we move beyond the single aggregate scores reported in prior work to systematically evaluate 12 frontier LLMs on 100 legal cases ranging from 32K to 512K tokens, primarily from 2025. Our detailed analysis reveals Gemini 2.5 Pro, Claude Sonnet 4, and Gemini 2.5 Flash achieve the best performance (around 50 $S_{\text{GAVEL-REF}}$), showing the difficulty of the task. These top models show consistent patterns: they succeed on simple checklist items (e.g., filing date) but struggle on multi-value or rare ones such as settlements and monitor reports. As LLMs keep improving and may eventually surpass human summaries, we also explore checklist extraction directly from case documents. We experiment with three different methods: end-to-end with long-context LLM, chunk-by-chunk extraction, and our newly developed autonomous agent scaffold, GAVEL-AGENT. Our results show strong potential for the agent approach in long-context processing: compared to the best GPT-4.1 end-to-end setup, Gavel-Agent with Qwen3 reduces token usage by 36% while achieving competitive performance (only 7% lower in $S_{\text{checklist}}$). We will release our code and annotations publicly to facilitate future research on long-context legal summarization.

1 INTRODUCTION

Large language models (LLMs) (Brown et al., 2020; Achiam et al., 2023) are now widely adopted across various industries and professions. The legal sector has been particularly active (Frankenreiter & Nyarko, 2022; Ziffer, 2023), with startups such as Harvey building AI for lawyers. Among legal applications, court document summarization stands out as both practically important and technically challenging. A single litigation case can easily involve dozens of court documents, including complaints, orders, and rulings, with a combined length exceeding 100,000 tokens, roughly equivalent to 80 news articles or a 300-page novel. Unlike news summarization, where lead sentences often suffice (Narayan et al., 2018; Liu & Lapata, 2019), or fiction books, where events can be summarized sequentially (Chang et al., 2024), legal cases require tracking interconnected arguments across multiple documents. It requires maintaining exact chronology, preserving relationships between parties, claims, and rulings, and ensuring that cross-references between filings remain accurate. Moreover, a collection of expert-written case summaries is available (Shen et al., 2022) to serve as a gold standard for this task. The combination of these factors makes legal summarization an ideal testbed for assessing LLMs' long-context capabilities; meanwhile, it also calls for more reliable and comprehensive evaluation methodologies than those currently in use.

To evaluate summarization, researchers have moved beyond traditional n-gram metrics such as ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002), developing checklist-based methods with LLM-as-judge (Min et al., 2023; Pereira et al., 2024; Lee et al., 2024; Lin et al., 2025). The most


14 documents
Docket
Complaint
Motion
Motion
Motion
Motion
Gemini 2.5 Pro
492 words
Human Summary
1,277 words
LLM
LLM
Checklist Evaluation
Extract checklist items
Filing Date: February 28, 2023
Parties (3 values): Plaintiff: The United States of America
Defendant: A neoprene manufacturer ...
Defendant: the manufacturer's landlord ...
Remedy Sought (2 values):
Injunction ... the manufacturer
Preliminary injunction ...
Dates of All Decrees (1 value):
March 10, 2025: order of dismissal.
22 more checklist items
Filing Date: February 28, 2023
Parties (3 values): Plaintiff: United States government
Defendant: Denka Performance ...
Defendant: Dupont ...
Remedy Sought (4 values):
Injunctive relief directing Denka ...
Injunctive relief ordering Dupont ...
Denka's counterclaim ...
Preliminary injunction ...
Dates of All Decrees (4 values):
August 30, 2023: ...
January 10, 2025: ...
July 2024: ...
March 10, 2025: the case was dismissed.
Compare checklist items
LLM
Equals
LLM summary doesn't contain the names
LLM summary misses 2 remedies including one sought by the defendants
LLM summary only covers 1 out of 4 decrees
19 applicable items:
2 Equal
2 Contains
2 Not Equal
7 Partial Match
2 Missing
4 Hallucination
$S_{\text{checklist}}$: 38 / 100
Residual Facts Evaluation
Get text that are not covered by the checklist items by values and supporting text of each extracted item
From Model Summary:
From Human Summary:
Extract residual facts
LLM
And compare them
4 residual facts
12 residual facts
On May 16, 2024, the EPA published a final rule regulating ...
Following the EPA rule, the district-court case was held in abeyance.
...
$F_1 = 0.25$
rescale
S_{residual}: 25 / 100
Writing Style Evaluation
Readability & Jargon Level: 3
Narrative Order: 4
Sentence Structure & Voice: 3
Formatting & Layout: 5
Citation & Reference Style: 4
1-5 Likert Scale on Similarity
Avg: 3.8
rescale
S_{style}: 70 / 100

On February 28, 2023, the United States of America, on behalf of the Environmental Protection Agency (EPA), filed a lawsuit in the U.S. District Court for the Eastern District of Louisiana against a neoprene manufacturer and its landlord. The government was represented by the Department of Justice and EPA counsel, while the defendants retained private counsel.

The lawsuit was brought under the "imminent and substantial endangerment" provision of the Clean Air Act, 42 U.S.C. § 7603. The United States alleged that the manufacturer's facility in LaPlace, Louisiana–the only one of its kind in the country–was emitting the chemical chloroprene at levels that posed an unacceptably high cancer risk to the surrounding community, ...

This case challenges the carcinogenic chloroprene emissions from Denka Performance Elastomer, LLC's (Denka's) neoprene manufacturing facility at the Pontchartrain Works Site in St. John the Baptist Parish, Louisiana, as being unacceptably high under EPA regulations. Denka's facility produced neoprene, a synthetic rubber, emitting chloroprene–a chemical classified by the EPA in 2010 as likely carcinogenic to humans. Air monitoring near the facility reportedly detected chloroprene concentrations up to 14 times the EPA's recommended levels, posing a risk to the majority-Black population surrounding the facility.

On February 28, 2023, the United States government filed a lawsuit in the U.S. District Court for the Eastern District of Louisiana. Plaintiff sued Denka Performance Elastomer LLC, the owner and operator of a neoprene manufacturing facility, and DuPont Specialty Products, the owner of the land on which the neoprene manufacturing facility was located and leased to Defendant Denka. Plaintiff sued Defendant Denka under Section 303 of the Clean Air Act, 42 U.S.C. § 7603, alleging that Defendant Denka's chloroprene ...

Figure 1: Example of evaluating a Gemini 2.5 Pro summary with GAVEL-REF, which contains: checklist evaluation supporting both string-wise and list-wise comparisons, residual fact evaluation, and writing-style evaluation. An interesting finding is that many modern LLMs tend to omits specific names of people or organizations—in this case, the defendant companies; and in other cases even the U.S. president's name. Light green background indicates matched values.

relevant recent work is ExpertLongBench (Ruan et al., 2025), which includes legal summarization in its benchmark. They ask legal experts to define 26 checklist items commonly found in legal case summaries (e.g., filing date, remedy sought, decrees), and an LLM is used to extract these items from both human- and model-generated summaries for item-by-item comparison. This marks an important step toward structured and interpretable evaluation, but the approach still has two key limitations: (i) many checklist items (e.g., remedy sought) may contain multiple distinct values (see Figure 1), yet existing method treats them as a single text block, making it difficult to capture partial matches. (ii) the evaluation is restricted to predefined checklist items, overlooking additional useful content outside the checklist and other qualities such as readability or formatting. Furthermore, ExpertLongBench and other existing benchmarks (Yen et al., 2024; Ruan et al., 2025) are built to evaluate LLMs across many tasks, with legal summarization as one of them. They provide valuable benchmarking of these models, but they naturally do not aim to offer detailed analysis of how modern LLMs perform on legal summarization specifically—for example, which checklist items models systematically struggle with or whether they capture non-checklist information that human experts often include. Finally, as LLMs continue to advance, they may surpass human-written summaries. This motivates deriving checklists directly from case documents to reduce reliance on human references while enabling test-time feedback. However, it is unclear from existing work whether current LLMs or agent-based methods can effectively handle this long-context extraction task.

In this paper, we address all three gaps. Firstly, we introduce GAVEL-REF (see Figure 1), which improves checklist evaluation by enabling list-wise comparison, and we further extend it with assessments of residual facts (information beyond the 26 checklist items) and writing style. We compare GAVEL-REF, with different LLMs as its backbone, against human annotators who perform the same task. Specifically, we collect 5,442 item-level annotations on 40 long summaries (averaging 1,130 words each), 450 checklist comparison judgments, and 375 style similarity ratings, totaling 150 hours of human effort. Our results show that GAVEL-REF using open-source GPT-oss 20B (Agarwal et al., 2025) and Qwen3 (Yang et al., 2025) models achieves performance comparable to GPT-5, demonstrating that large-scale automatic evaluation can be both reliable and cost-effective.

Secondly, using GAVEL-REF, we evaluate 12 LLMs, including proprietary models (GPT-5 and Gemini 2.5) and open-source models (GPT-oss and Qwen3), on 100 cases spanning 32K to 512K tokens, far beyond the 128K limit of prior work. To reduce data contamination, 83% of cases are new from 2025 and likely unseen by the models. Our main findings are: (i) Gemini 2.5 Pro, Claude Sonnet 4, and Gemini 2.5 Flash achieve the best summaries with $S_{\text{GAVEL-REF}}$ score of 50 out of 100, underscoring the difficulty of long-context legal summarization. (ii) Proprietary models outperform open-source ones at the 30B scale, with open-source models such as Gemma3 (Team et al., 2025) and Qwen3 degrading more drastically as case length increases. (iii) GPT-4.1 best captures residual facts while GPT-5 tends to produce checklist-like and verbose summaries even when prompted for narrative style, while Claude and Gemini models most closely match human style. (iv) Top models handle single-value items well, multi-value items less reliably, and struggle most with related cases and monitoring reports.

Thirdly, for extracting checklists directly from case documents, beyond standard approaches such as feeding all documents into a long-context LLM or chunking them and extracting items iteratively, we develop a novel agent scaffold, GAVEL-AGENT. It equips LLMs with six tools for autonomously navigating documents and locating checklist items, emulating how humans process case documents. Our experiments show that end-to-end extraction with GPT-4.1 achieves the best overall performance, with GAVEL-AGENT using Qwen3 performing very closely behind. The advantage of GAVEL-AGENT is efficiency: it uses 36% fewer tokens than the GPT-4.1 end-to-end setup and 59% fewer than the chunk-by-chunk approach, highlighting the strong potential of agents for long-context tasks. Compared to extracting from summaries, checklist extraction from full documents still lags significantly, pointing to future work on long-context LLMs and long-horizon agents.

In summary, our contributions are as follows:

1. We introduce GAVEL-REF, a reference-based evaluation framework for legal summarization that provides a comprehensive assessment via checklist, residual fact, and writing style evaluation.
2. Using GAVEL-REF, we systematically evaluate 12 frontier LLMs across different case lengths and reveal their gaps in capturing complex legal checklist items with a detailed analysis.
3. We explore checklist extraction from case documents using three different approaches: end-to-end, chunk-by-chunk, and GAVEL-AGENT—our autonomous agent scaffold.

2 GAVEL-REF—A REFERENCE-BASED EVALUATION FRAMEWORK

We introduce GAVEL-REF (Fig. 1), an automatic, reference-based evaluation framework for legal summarization with three complementary components. First, *checklist evaluation* extracts values and supporting text for 26 items(e.g., filing date, parties, decrees). Second, *residual facts evaluation* captures and scores content beyond the checklist. Third, *writing style evaluation* compares model summaries' similarity to human references across five aspects. Prompts are in App. G.

2.1 METHOD DESCRIPTION

Checklist Evaluation. ExpertLongBench (Ruan et al., 2025) presents a checklist-based evaluation framework for long-form generation, where legal experts create a checklist of 26 key items for legal summaries. For each item c_i, an LLM extracts the corresponding information $H(c_i)$ from the model summary and $R(c_i)$ from the reference, then determines containment relationships between them. While this provides a solid foundation, we identify limitations and improve it as follows:

Improvement 1: Multi-value extraction with supporting text. We find that checklist items contain multiple values 76% of the time (e.g., several filings or factual bases in a case). However, prior method extracts all information as a single text block and performs a binary comparison. This misses partial overlaps—for example, five filings vs. five different filings with three overlaps is scored the same as a total mismatch.

To address this limitation, we restructure extraction so that each checklist item c_i yields a list of values with supporting text: $H(c_i) = \{(v_{i,1}, s_{i,1}), (v_{i,2}, s_{i,2}), \ldots, (v_{i,n}, s_{i,n})\}$, where $v_{i,j}$ is the j-th extracted value for checklist item c_i, and $s_{i,j}$ is a set of verbatim snippets grounding it. Supporting text not only justifies values but also helps us later identify residual facts that fall outside the checklist. For comparison, single-value items are judged by an LLM as equal, A contains B, B contains A, or different, while multi-value items use element-wise matching to identify overlaps and uniques.

Improvement 2: Score aggregation. When some checklist item doesn't exist in the case documents, both the model and human naturally won't include it in their summaries. However, the original method counts it as a correct match. This inflates the denominator and reduces the penalty for actual errors. As non-applicable items dilute the score calculation, errors like hallucinations or omissions of key items have less impact on the final score.

To address this issue, we compute scores based only on applicable items, defined as those present in at least one summary. The final score is: $S_{\text{checklist}} = \frac{100}{|A|} \sum_{c_i \in A} m_i$, where A is the set of applicable checklist items, and the matching score m_i is defined as:

$$m_i = \begin{cases} \begin{cases} 1 & \text{if } H(c_i) = R(c_i) \\ 0.5 & \text{if } H(c_i) \subset R(c_i) \text{ or } H(c_i) \supset R(c_i) \\ 0 & \text{otherwise} \end{cases} & \text{if single-value} \\ F_1(H(c_i), R(c_i)) & \text{if multi-value} \end{cases} \tag{1}$$

For single-value items, we assign full points for equality, half points for containment, and zero otherwise. For multi-value items, we use F_1 as the matching score.

Residual Facts Evaluation. While the checklist captures essential case information, summaries sometimes include details beyond these 26 items. To evaluate this additional content, we first identify text segments not covered by the checklist. We use two-stage matching to precisely identify uncovered text: first against the extracted values alone, then against their supporting sentences if unmatched. This prevents over-coverage—such as when a filing date's support text also contains other legal facts. We then use an LLM to extract atomic facts (termed "residual facts") from these uncovered segments and evaluate them using the same list-wise comparison method as in our checklist evaluation. The resulting F_1 score (scaled to 0-100) is the S_{residual}.

Writing Style Evaluation. Beyond content, we measure how closely model summaries match human ones in writing style. We emphasize similarity over quality, as quality is subjective (e.g., preference for narratives vs. bullet points). Five aspects are rated on a 1–5 Likert scale (1 = completely different, 5 = identical): Readability & Jargon Level, Narrative Order, Sentence Structure & Voice, Formatting & Layout, Citation & Reference Style. We average these scores, subtract 1, and multiply by 25 to obtain S_{style} on a 0-100 scale. See Appendix C for definitions of each aspect.

2.2 The Overall Gavel-Ref Score

To combine all three components into a final score for benchmarking LLMs or use as a reward signal, we compute a weighted linear combination:

$$S_{\text{GAVEL-REF}} = (1 - r) \cdot \alpha \cdot S_{\text{checklist}} + r \cdot \alpha \cdot S_{\text{residual}} + (1 - \alpha) \cdot S_{\text{style}} \tag{2}$$

where α controls the balance between content and style, and r is the proportion of residual content in the reference summary (total residual text spans length divided by summary length). This dynamically weights $S_{\text{checklist}}$ and S_{residual} based on their relative importance in each summary—more residual content increases the weight on S_{residual}. We set α as 0.9 throughout our paper.

2.3 Meta-Evaluation of Gavel-Ref

To validate that Gavel-Ref accurately captures summary quality, we recruit four in-house annotators to perform the same evaluation tasks as the LLM—extracting checklist items, comparing checklist item values, and rating writing style similarity—then measure the agreement between LLM and human annotations.

Collecting Human Annotations. To evaluate LLMs' ability to *extract checklist items*, we annotated 40 long case summaries (avg. 1,130 words) to stress-test the models: if the LLM can accurately extract checklist items from these longer summaries, it should perform at least as well on the shorter ones used in the main model evaluation. Since extracting all 26 checklist items from scratch is time-consuming, annotators start from GPT-5's extractions. Using our paragraph-by-paragraph review interface modified from Thresh (Heineman et al., 2023), annotators add missing values, correct extractions and supporting text, or delete incorrect values. Each summary annotation takes approximately one hour. Figures 13 to 22 in the Appendix show an example of our annotations on a case summary, covering all 26 checklist items. In total, we collect 70 summary-level annotations covering 5,442 item-level annotations, where the ten longest summaries (averaging 1,695 words) receive triple annotations, with adjudication by a fourth annotator. The remaining 30 summaries receive single annotations. To evaluate LLMs' ability to *compare checklist values*, annotators assess 150 item pairs from model and reference summaries (100 multi-value, 50 single-value), drawn from diverse LLMs for generalizability. For single-value pairs, they perform 4-class classification: equal, A contains B, B contains A, or different. For multi-value pairs, they match elements from list A to list B. Annotations are aggregated by majority vote: for single-value items, we take the class with $\geq$ two votes (no cases had all three labels differ); for multi-value items, we keep matches identified by $\geq$ two annotators. To evaluate LLM's ability to *rate writing style similarity*, we annotate 25 model-reference summary pairs. Annotators rate similarity across five style aspects using 1-5 Likert scales, with three annotations per pair. Final scores are the median across annotators.

All annotators are paid $18 USD per hour, with a total cost of $3K USD. Appendix D provides training details, inter-annotator agreement results, and screenshots of the annotation interfaces.

Metrics. For *checklist comparison*, we use accuracy for single-value items (4-class classification) and matching-pairs F1 for multi-value items, which measures how accurately the LLM identifies correct matches between two lists. The best comparison model is then used to evaluate *checklist extraction*, computing Schecklist against human-extracted checklist from the same summary. We also compute word-level coverage agreement on supporting text, measuring how often model and human agree on whether words are covered by checklist items or are residual. For *writing style rating*, we report Cohen's Kappa for LLM-human agreement.

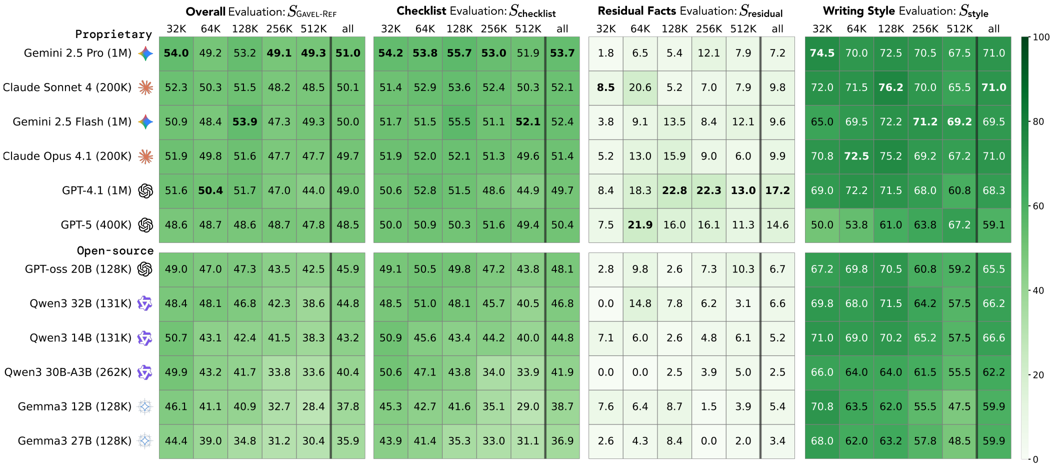
Results. We select models based on two criteria: state-of-the-art performance and open-source availability. We prioritize open-source models for cost-efficient large-scale evaluation in Section 3. We evaluate five LLMs: GPT-5 and four open-source models—Qwen3 32B, Qwen3 30B-A3B, GPT-oss 20B, and Gemma3 27B. Table 1 presents the results. GPT-5 performs best at checklist extraction, with GPT-oss 20B second overall and showing much higher coverage than the other open-source models. Reasoning models perform better than Gemma3 27B on this task. However, Gemma3 27B outperforms all reasoning models on single string comparison and achieves comparable performance on list-wise comparison. GPT-oss 20B achieves the best alignment with human ratings of writing style. Based on these results, we use GPT-oss 20B for checklist extraction and style rating, and Gemma3 27B for checklist comparison in Section 3 when evaluating LLM summaries.

	Checklist Extraction		Checklist Comparison		Style
Model	$S_{\text{checklist}}$	Coverage	Single	Multi	Rating
GPT-5	**68.2**	**92.9%**	0.567	*0.847*	*0.115*
GPT-oss 20B	64.4	*83.7%*	0.567	0.801	**0.157**
Gemma3 27B	54.1	75.3 %	**0.740**	0.841	0.091
Qwen3 32B	*65.5*	66.0%	0.600	0.820	0.084
Qwen3 30B-A3B	63.3	63.0%	*0.700*	**0.854**	-0.011

Table 1: Meta-evaluation results of five models in GAVEL-REF: Checklist Extraction ($S_{\text{checklist}}$ and word-level coverage agreement), Checklist Comparison (accuracy for single-value, matching F_1 for multi-value), and Writing Style Rating (Cohen's κ). **Bold**: best, *italic*: second best.

3 EVALUATION OF LLM LEGAL SUMMARIZATION WITH GAVEL-REF

Prior work (Yen et al., 2024; Ruan et al., 2025) have evaluated LLM legal summarization on legal cases up to 128K that are before 2024. As the latest LLMs now handle 1M tokens and have pretrained knowledge up to 2025, in this work, we want to shed light on how these modern models perform on much longer context using 2025 legal cases beyond their training cutoffs. With GAVEL-REF, we evaluate 12 LLMs that span both proprietary and open-source models across 5 different case length scales: 32K, 64K, 128K, 256K, 512K tokens (measured by the GPT-4o tokenizer). For each scale, we select 20 cases whose token counts fall within ±20% of the target length. Of the 100 cases, 83 are filed in 2025 (using the filing date of the first docket entry). The remaining 17 cases (14 in the 512K bin and 3 in the 32K bin) are from earlier years due to limited availability—especially for the 512K bin. At the time of writing (7-8 months into 2025), very few cases have accumulated enough documents to reach 512K tokens; on average, cases in this bin take about 1.5 years to reach that length. Since the models have varying context limits and some cases exceed these limits, we truncate by proportionally removing tokens from the end of each document, following prior work.



	Overall Evaluation: $S_{\text{GAVEL-REF}}$						Checklist Evaluation: $S_{\text{checklist}}$						Residual Facts Evaluation: S_{residual}						Writing Style Evaluation: S_{style}					
	32K	64K	128K	256K	512K	all	32K	64K	128K	256K	512K	all	32K	64K	128K	256K	512K	all	32K	64K	128K	256K	512K	all
Proprietary																								
Gemini 2.5 Pro (1M)	**54.0**	49.2	53.2	**49.1**	**49.3**	**51.0**	**54.2**	**53.8**	**55.7**	**53.0**	51.9	**53.7**	1.8	6.5	5.4	12.1	7.9	7.2	**74.5**	70.0	72.5	70.5	67.5	71.0
Claude Sonnet 4 (200K)	52.3	50.3	51.5	48.2	48.5	50.1	51.4	52.9	53.6	52.4	50.3	52.1	**8.5**	20.6	5.2	7.0	7.9	9.8	72.0	71.5	**76.2**	70.0	65.5	**71.0**
Gemini 2.5 Flash (1M)	50.9	48.4	**53.9**	47.3	49.3	50.0	51.7	51.5	55.5	51.1	**52.1**	52.4	3.8	9.1	13.5	8.4	12.1	9.6	65.0	69.5	72.2	**71.2**	**69.2**	69.5
Claude Opus 4.1 (200K)	51.9	49.8	51.6	47.7	47.7	49.7	51.9	52.0	52.1	51.3	49.6	51.4	5.2	13.0	15.9	9.0	6.0	9.9	70.8	**72.5**	75.2	69.2	67.2	71.0
GPT-4.1 (1M)	51.6	**50.4**	51.7	47.0	44.0	49.0	50.6	52.8	51.5	48.6	44.9	49.7	8.4	18.3	**22.8**	**22.3**	**13.0**	**17.2**	69.0	72.2	71.5	68.0	60.8	68.3
GPT-5 (400K)	48.6	48.7	48.6	48.7	47.8	48.5	50.0	50.9	50.3	51.6	49.4	50.4	7.5	**21.9**	16.0	16.1	11.3	14.6	50.0	53.8	61.0	63.8	67.2	59.1
Open-source																								
GPT-oss 20B (128K)	49.0	47.0	47.3	43.5	42.5	45.9	49.1	50.5	49.8	47.2	43.8	48.1	2.8	9.8	2.6	7.3	10.3	6.7	67.2	69.8	70.5	60.8	59.2	65.5
Qwen3 32B (131K)	48.4	48.1	46.8	42.3	38.6	44.8	48.5	51.0	48.1	45.7	40.5	46.8	0.0	14.8	7.8	6.2	3.1	6.6	69.8	68.0	71.5	64.2	57.5	66.2
Qwen3 14B (131K)	50.7	43.1	42.4	41.5	38.3	43.2	50.9	45.6	43.4	44.2	40.0	44.8	7.1	6.0	2.6	4.8	6.1	5.2	71.0	69.0	70.2	65.2	57.5	66.6
Qwen3 30B-A3B (262K)	49.9	43.2	41.7	33.8	33.6	40.4	50.6	47.1	43.8	34.0	33.9	41.9	0.0	0.0	2.5	3.9	5.0	2.5	66.0	64.0	64.0	61.5	55.5	62.2
Gemma3 12B (128K)	46.1	41.1	40.9	32.7	28.4	37.8	45.3	42.7	41.6	35.1	29.0	38.7	7.6	6.4	8.7	1.5	3.9	5.4	70.8	63.5	62.0	55.5	47.5	59.9
Gemma3 27B (128K)	44.4	39.0	34.8	31.2	30.4	35.9	43.9	41.4	35.3	33.0	31.1	36.9	2.6	4.3	8.4	0.0	2.0	3.4	68.0	62.0	63.2	57.8	48.5	59.9

Figure 2: Benchmarking results of 12 LLMs on long-context legal summarization with our GAVEL-REF framework across case lengths from 32K to 512K tokens. Models are ordered by $S_{\text{GAVEL-REF}}$ on all cases. Gemini 2.5 Pro leads, with all top six positions held by proprietary models.

3.1 BENCHMARKING RESULTS FOR 12 MODELS

Figure 2 shows GAVEL-REF evaluation results for 12 models across different case length bins. Figure 6 in the Appendix additionally shows the summary length of each model in each length bin, compared to human summary length.

Gemini 2.5 Pro, Claude Sonnet 4, and Gemini 2.5 Flash are the top three models. Proprietary models consistently outperform open-source ones by a clear margin. Overall, Gemini 2.5 Pro achieves the best performance with an $S_{\text{GAVEL-REF}}$ of 51.0, while the best open-source model, GPT-oss 20B, reaches 45.9. Interestingly, GPT-5 is the weakest among the proprietary models, largely due to its overly verbose summaries, which we analyze in more detail in the paragraphs below. Within the Claude family, Sonnet 4 slightly outperforms Opus 4.1. To understand which checklist items drive this gap, we present checklist item–level performance for each LLM in Figures 10–12 in the Appendix. We find that Sonnet 4 is stronger in identifying items such as Cause of action, Class action vs. individual, and Remedy sought than Opus 4.1.

All models degrade as case length increases, with larger drops for open-source models. We observe a consistent pattern: $S_{\text{GAVEL-REF}}$ decreases as case length grows, and models perform worst on the 256K and 512K bins. Even though models like Gemini 2.5 Pro, Gemini 2.5 Flash, and GPT-4.1 support a 1M-token context window, they still show noticeable drops on long cases—for example, Gemini 2.5 Pro is 4.7 points lower on 512K than on 32K cases, and GPT-4.1 drops by 7.6 points. Open-source models degrade even more on 256K and 512K cases, which is expected since they do not support such long contexts, and truncation of the case documents causes substantial information loss. These results call for scaffolded agents for long-context legal summarization.

GPT-4.1 performs best on residual facts evaluation, with GPT-5 close behind. Both models tend to capture more non-checklist details than other models. On average, the residual ratio r (the proportion of residual content in the whole summary, Eq. 2) is 18.7% for GPT-4.1 and 18.4% for GPT-5. These are the only two models that exceed the human residual ratio of 11.1%; the next highest model, Claude Sonnet 4, is only 7.3%. As a result, GPT-4.1 and GPT-5 obtain the highest S_{residual} of 17.2 and 14.6, respectively. However, these values are still below 20, indicating that the overlap between human residual facts and the residual facts captured by the models remains limited.

Surprisingly, GPT-5 has the lowest writing-style rating, while Gemini and Claude models have the most human-like style. Claude Opus 4.1, Sonnet 4, and Gemini 2.5 Pro all achieve S_{style} of 71.0, whereas GPT-5 scores lowest at 59.1. As illustrated in Figure 9, GPT-5 often ignores the instruction to write in narrative form, instead producing sectioned summaries organized by checklist items, and tends to be very verbose—sometimes close to 1,000 words when the corresponding human summary is around 700 words. All models perform best on 64K–128K cases in terms of style similarity. On longer cases (256K–512K), every model's writing becomes less human-like, with similar drops across the board. From Figure 6, we see that in the 256K and 512K bins human summaries are

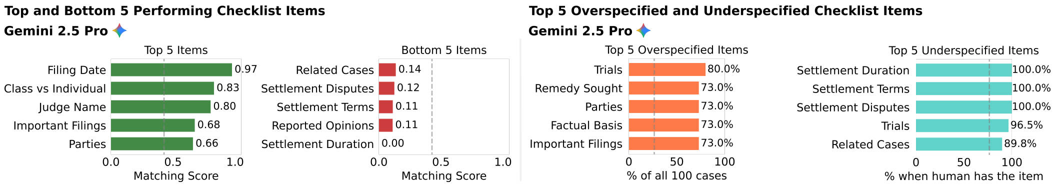


Figure 4: Gemini 2.5 Flash performance breakdown: top/bottom 5 checklist items by matching score and most frequently over/under-specified items. Overspecification measured as frequency across all 100 cases; underspecification as frequency among cases where human summary includes that item. Dashed lines are medians: 0.49 matching score, 59% overspecification, 70% underspecification.

around 1,200 words, while proprietary models (excluding GPT-5) typically produce summaries of 500–800 words. Open-source models are even more concise, usually under 400 words, and weaker models such as Gemma often stay below 300 words across all length bins.

3.2 How Top Models Handle Different Checklist Information

Figure 3 shows performance of the top five models across nine checklist groups, using the matching score m_i (Eq. 1). All models follow a similar pattern. **They are good at extracting basic case information, legal foundations, and judge details**, scoring above 0.6. This makes sense as these groups contain mostly single-value items like filing date, cause of action, type of counsel, and judge name. **Performance drops noticeably for multi-value items.** Court rulings, decrees, and factual basis (context) prove more challenging, with scores around 0.4-0.5. Models must track multiple related pieces of information scattered across lengthy documents and determine which ones are important enough to include. **The models struggle most with related cases and settlements,** scoring below 0.2. The items in these groups appear rarely in the cases.

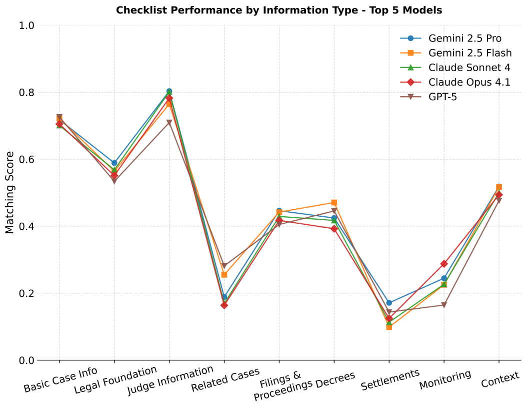


Figure 3: Top-5 LLMs' performance across checklist groups, struggling the most on rare items such as related cases and settlements.

3.3 Dissecting the Top Performer: Item-Level Analysis

Figure 4 analyzes Gemini 2.5 Pro's item-level performance, showing its top and bottom 5 checklist items plus consistently over- and under-specified items (see Appendix Figure 7 for top-3 models).

Single-value items are Gemini's strength, while settlement details are its blind spots. Filing date leads with a near-perfect matching score of 0.97, followed by other straightforward items such as Class action vs. Individual (0.83) and Judge name (0.80). For the next-best items, Important Filings and Parties, the scores fall below 0.7, and the median matching score across all 26 items is 0.43. In contrast, Gemini struggles dramatically with settlement-related information—scoring just 0.12, 0.11, and 0.00 on the three settlement items—while Related Cases and Reported Opinions are also among the weakest-performing items.

Gemini 2.5 Flash tends to overspecify and underspecify checklist items with multiple values in its summaries. All of the top five over-specified and under-specified items are multi-value items, with Trials appearing in both lists. This suggests that when multiple values are possible, the model has difficulty matching human judgments about which details to include. Settlement Duration, Settlement Terms, and Settlement Disputes are under-specified 100% of the time. Overall, the model is much more prone to under-specification than over-specification: the median overspecification rate is 26.5%, whereas the median underspecification rate is 76.5%.

4 EXTRACTING CHECKLIST FROM CASE DOCUMENTS

While reference-based evaluation effectively benchmarks summarization models, it requires hours of legal expert time per case to create human summaries, which cannot serve as a long-term gold standard once LLMs begin to surpass humans. Directly extracting checklists from case documents removes this dependency, enabling scalable evaluation, testing of superhuman models, and grounded suggestions during inference. To this end, we experiment with three methods: end-to-end extraction with long-context LLMs, processing the case documents chunk by chunk, and GAVEL-AGENT—an autonomous agent framework we develop to test whether LLMs can efficiently extract information by strategically searching and skimming rather than reading every word.

4.1 METHODS

End-to-end. We concatenate all case documents in chronological order and feed them to long-context LLMs. Instead of extracting all 26 checklist items at once, we query each item individually, which gives more accurate results.

Chunk-by-chunk. We split each document into 16K-token chunks, long enough to capture most documents while fitting within modern LLM context windows (32K+). At each step, the model receives the chunk text and current checklist state, then outputs an updated state—retaining existing values or adding new ones. Like end-to-end, we process documents chronologically and extract all 26 items. This mirrors multi-agent long-context methods (Zhang et al., 2024; Zhao et al., 2024), which segment text and process chunks independently.

GAVEL-AGENT. Unlike end-to-end or chunk-by-chunk methods that make models to read everything, human experts strategically search and skim for relevant information. To mimic this, we develop GAVEL-AGENT, an agent scaffold that lets LLMs navigate documents and extract checklist items autonomously. GAVEL-AGENT provides the LLM with six tools such as read a document, run regex searches across documents, and update checklist items. At each step, the model chooses a tool or issues a stop action based on the current state and history. Standard scaffolds append each tool call and response to agent's context. While working for short tasks, this approach breaks down in long cases (256K+ tokens, 50+ calls), where the context quickly balloons and the model must track information across an increasingly unwieldy history. Instead, GAVEL-AGENT refreshes the state after each tool call, giving LLM a clean snapshot including documents explored state, recent action details, etc. GAVEL-AGENT is fully customizable: users can define any checklist items, making it easy to transfer to domains like biomedical or financial extraction.

Tools. The following are the definitions of the six tools in GAVEL-AGENT:

- `list_documents()`: Returns all available documents with their metadata such as document type and token count. It is used to provide an initial catalog of the case.
- `read_document(doc_name, start_token, end_token)`: Reads a specific token range from a document, with a maximum of 10,000 tokens per call.
- `search_document_regex(pattern, doc_name/doc_names, top_k, context_tokens)`: Searches one, multiple or all documents using regex patterns, returning the top-k matches with surrounding context (100-1000 tokens).
- `get_checklist(item/items)`: Retrieves extracted values for specified checklist items.
- `append_checklist(patch)`: Adds new values for specific checklist items, supporting multiple values per item with required evidence (verbatim text, source document, and location).
- `update_checklist(patch)`: Replaces all values for specified checklist items, used for corrections or marking items as ”Not Applicable” when no relevant information exists.

Both `append_checklist` and `update_checklist` use a `patch` structure that supports batch operations. Each patch contains an array of checklist keys to update, where each key maps to an array of extracted values, and every value includes (1) the value itself and (2) an array of supporting evidence (verbatim text, source document, and location). This structure ensures traceability from extracted information back to source documents.

Context Management. At each step, the LLM is given a system prompt high-level task instruction and tool descriptions, and a user prompt that contains user instruction (e.g., “Extract all 26 checklist items”), the checklist definitions of the items to extract, a document catalog showing which parts have been explored, a summary of what has been extracted so far, and the recent action history. For action history, we maintain up to 100 tool calls: the five most recent include full responses (e.g., full text from `read_document`), while the other 95 are compressed to the tool name and brief outcome (e.g., “read 3,000 tokens”, “updated filing date”). This gives the model enough awareness to avoid repeating actions while keeping the prompt compact.

4.2 Implementation Details

Model Selection. For end-to-end extraction, we use GPT-4.1 with its 1M-token context. For chunk-by-chunk extraction, we test three open-source reasoning models: GPT-oss 20B, Qwen3 32B, and Qwen3 30B-A3B. For GavelAgent, we use Qwen3 30B-A3B and GPT-oss 20B, as both support 128K+ context natively, sufficient for context management.

Gavel-Agent Configurations. It is unclear whether agents perform better extracting multiple checklist items together—potentially using each document read more efficiently—or focusing on single items for higher accuracy. To study this trade-off, we test three setups: (1) one agent extracting all 26 items; (2) 9 agents for grouped items (e.g., filing date, parties, and counsel under “Basic Case Information”); (3) 26 agents, each handling a single item. See App. B for full checklist definitions.

4.3 Meta-Evaluation

Following the evaluation of Gavel-Ref in Section 2.3, we evaluate extraction quality on 20 long cases. We use Gemma3 27B to compare each method’s extracted checklist against the human-created checklist from the summary, computing the $S_{\text{checklist}}$ score. We also measure token usage (input and output) as efficiency.

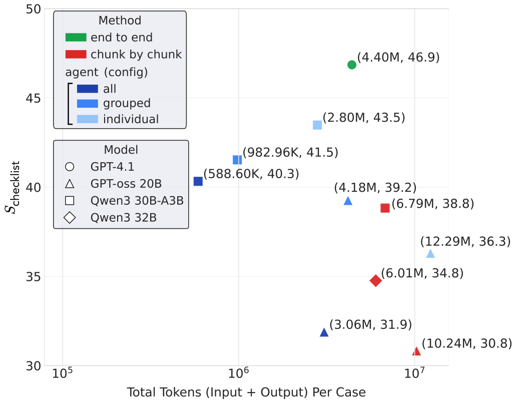


Figure 5: $S_{\text{checklist}}$ versus total token usage for different methods extracting from case documents.

Results. Figure 5 shows $S_{\text{checklist}}$ versus total token usage for each method (input and output token breakdowns are in Figure 8 in the Appendix.) End-to-end extraction with GPT-4.1 achieves the highest $S_{\text{checklist}}$ of 46.9 but uses 4.4M tokens. Gavel-Agent with 26 individual agents using Qwen3 30B-A3B achieves the second-best $S_{\text{checklist}}$ of 43.5 while using only 2.8M tokens. This is 36% fewer tokens than end-to-end with GPT-4.1 and 59% fewer than the chunk-by-chunk method with the same Qwen3 model. Within the Gavel-Agent configurations, we see a clear quality-cost trade off. A single agent extracting all 26 items is the most token-efficient but provides the lowest $S_{\text{checklist}}$. For Qwen3 30B-A3B, the 26-agent configuration achieves the best performance, and the grouped configuration lies in between on both quality and token usage. This shows that, in our setting, agents work better when they focus on fewer items at a time; in the future, being able to reliably handle multiple items per read could unlock further token savings. The best chunk-by-chunk performance is 38.8 with Qwen3 30B-A3B, which is much lower than end-to-end and Gavel-Agent. Overall, these results show strong potential for autonomous agents to process long-context inputs, delivering substantially better efficiency while achieving competitive top-level performance. Notably, all document extraction methods fall well below the 68.2 achieved by GPT-5 extracting from human summaries in Gavel-Ref, showing significant headroom for improving both long-context models and long-horizon agents.

5 Related Work

Legal Summarization. Several datasets exist for this task. Shukla et al. (2022) release Indian and UK Supreme Court cases with human-written summaries, and Elaraby & Litman (2022) provide Canadian court opinions paired with expert summaries. Heddaya et al. (2024) collect U.S. Supreme

Court opinions with their official summaries. These resources focus on single-document summarization with inputs under 16K tokens. Multi-LexSum (Shen et al., 2022) and ExpertLongBench (Ruan et al., 2025) extend this to multi-document summaries using cases from the Civil Rights Litigation Clearinghouse (CRLC), a widely used platform that offers free access to U.S. civil rights cases. Following them, we also collect cases from CRLC, focusing on 2025 filings to reduce data contamination. To better evaluate long-context capability, we construct five length ranges (32K–512K tokens) and benchmark 12 state-of-the-art LLMs with our framework GAVEL-REF, which provides fine-grained analysis of their strengths and weaknesses in long-context legal summarization.

Checklist-based Evaluation. With modern LLMs, text evaluation has moved from n-gram metrics such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004) to LLM-based methods. One line of work (Min et al., 2023; Scirè et al., 2024) extracts atomic facts from the summaries, and verifies each fact's correctness. While precise, it is limited by inconsistent definitions of what constitutes an 'atomic' fact (Hu et al., 2024) and by poor scalability to long texts. Another line (Lee et al., 2024; Qin et al., 2024; Lin et al., 2025; Cook et al., 2024; Furuhashi et al., 2025) uses LLMs to generate task-specific rubrics and then evaluates responses against each rubric item. In domain-specific settings, human experts often design checklists that capture key information; for example, Arora et al. (2025) ask physicians to write rubrics for medical conversations. The most relevant work, ExpertLongBench (Ruan et al., 2025), introduces expert-designed checklists for 11 tasks, including 26 items for legal summarization (e.g., filing dates, court rulings). Building on this, we improve checklist extraction by requiring evidence for each item and introducing list-wise comparison. We further augment checklist evaluation with residual-fact and writing-style assessments to provide a complete picture of summary quality. Finally, we extend checklist extraction directly to case documents, reducing reliance on human summaries when evaluating future superhuman models.

LLM Agent Scaffolds. Modern LLM agents are designed as autonomous problem-solvers that plan actions and invoke tools in a multi-step loop for tasks such as web browsing (Gur et al., 2023), coding (Yang et al., 2024), or general-purpose reasoning. Several open-source scaffolds have been introduced (Xie et al., 2023; Wang et al., 2025; Lu et al., 2025; Qiu et al., 2025). For long-context processing, recent approaches segment documents into chunks or convert them into graph structures (Chen et al., 2023; Sun et al., 2024; Li et al., 2024; Zhao et al., 2024; Zhang et al., 2024), which we adopt as our chunk-by-chunk method. Inspired by how human experts read legal case documents—skimming titles, prioritizing files, and searching for keywords rather than reading everything exhaustively—we develop GAVEL-AGENT, an autonomous scaffold that equips models with six tools for navigating case documents. For context management, unlike the standard approach of continually appending tool calls and responses, we update a snapshot after each tool call and prompt the LLM with it. This design helps maintain an up-to-date state within context limits, especially when models issue 50+ tool calls in sequence, which would otherwise exhaust context quickly.

6 CONCLUSION

We present GAVEL-REF, a reference-based framework for evaluating long-context legal summarization that improves checklist-based evaluation with multi-value and support text extraction, and adds residual fact assessment and writing-style evaluation. In our systematic study of 12 frontier LLMs with GAVEL-REF on 2025 cases ranging from 32K to 512K tokens, we find that even the top models—Gemini 2.5 Pro, Claude Sonnet 4, Gemini 2.5 Flash—reach only about 50 $S_{\text{GAVEL-REF}}$, highlighting the difficulty of legal summarization. Our analysis reveals consistent patterns: models perform well on simple single-value items but struggle with multi-value and rare ones, showing key areas for improvement. To reduce reliance on human summaries, we also explore checklist extraction directly from case documents. Comparing end-to-end, chunk-by-chunk, and our proposed GAVEL-AGENT approach, we find that end-to-end extraction with GPT-4.1 achieves the best performance, while GAVEL-AGENT with Qwen3 comes very close and reduces token usage by 36–59%. Looking ahead, advancing long-context models and long-horizon agents for legal summarization and document-level extraction is key to making AI more effective in legal practice.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. BooookScore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7Ttk3RzDeu.

Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. Walking down the memory maze: Beyond context limit through interactive reading. *arXiv preprint arXiv:2310.05029*, 2023.

Jonathan Cook, Tim Rocktäschel, Jakob Foerster, Dennis Aumiller, and Alex Wang. Ticking all the boxes: Generated checklists improve llm evaluation and generation. *arXiv preprint arXiv:2410.03608*, 2024.

Mohamed Elaraby and Diane Litman. ArgLegalSumm: Improving abstractive summarization of legal documents with argument mining. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.540/.

Jens Frankenreiter and Julian Nyarko. Natural language processing in legal tech. *Legal Tech and the Future of Civil Justice (David Engstrom ed.) Forthcoming*, 2022.

Momoka Furuhashi, Kouta Nakayama, Takashi Kodama, and Saku Sugawara. Are checklists really useful for automatic evaluation of generative tasks? *arXiv preprint arXiv:2508.15218*, 2025.

Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. A real-world webagent with planning, long context understanding, and program synthesis. *arXiv preprint arXiv:2307.12856*, 2023.

Mourad Heddaya, Kyle MacMillan, Anup Malani, Hongyuan Mei, and Chenhao Tan. Casesumm: a large-scale dataset for long-context summarization from us supreme court opinions. *arXiv preprint arXiv:2501.00097*, 2024.

David Heineman, Yao Dou, and Wei Xu. Thresh: A unified, customizable and deployable platform for fine-grained text evaluation. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-demo.30/.

Qisheng Hu, Quanyu Long, and Wenya Wang. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? *arXiv preprint arXiv:2411.02400*, 2024.

Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, and Pilsung Kang. Checkeval: Robust evaluation framework using large language model via checklist. *CoRR*, 2024.

Shilong Li, Yancheng He, Hangyu Guo, Xingyuan Bu, Ge Bai, Jie Liu, Jiaheng Liu, Xingwei Qu, Yangguang Li, Wanli Ouyang, et al. Graphreader: Building graph-based agent to enhance long-context abilities of large language models. *arXiv preprint arXiv:2406.14550*, 2024.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking LLMs with challenging tasks from real users in the wild. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=MKEHCx25xp.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*, 2019.

Pan Lu, Bowen Chen, Sheng Liu, Rahul Thapa, Joseph Boen, and James Zou. Octotools: An agentic framework with extensible tools for complex reasoning. *arXiv preprint arXiv:2502.11271*, 2025.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.741/.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL https://aclanthology.org/D18-1206/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. URL https://aclanthology.org/P02-1040/.

Jayr Pereira, Andre Assumpcao, and Roberto Lotufo. Check-eval: A checklist-based approach for evaluating text quality. *arXiv preprint arXiv:2407.14467*, 2024.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. InFoBench: Evaluating instruction following ability in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.772/.

Jiahao Qiu, Xuan Qi, Tongcheng Zhang, Xinzhe Juan, Jiacheng Guo, Yifu Lu, Yimin Wang, Zixin Yao, Qihan Ren, Xun Jiang, et al. Alita: Generalist agent enabling scalable agentic reasoning with minimal predefinition and maximal self-evolution. *arXiv preprint arXiv:2505.20286*, 2025.

Jie Ruan, Inderjeet Nair, Shuyang Cao, Amy Liu, Sheza Munir, Micah Pollens-Dempsey, Tiffany Chiang, Lucy Kates, Nicholas David, Sihan Chen, et al. Expertlongbench: Benchmarking language models on expert-level long-form generation tasks with structured checklists. *arXiv preprint arXiv:2506.01241*, 2025.

Alessandro Scirè, Karim Ghonim, and Roberto Navigli. FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.841/.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *Advances in Neural Information Processing Systems*, 35:13158–13173, 2022.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. Legal case document summarization: Extractive and abstractive methods and their evaluation. *arXiv preprint arXiv:2210.07544*, 2022.

Simeng Sun, Yang Liu, Shuohang Wang, Dan Iter, Chenguang Zhu, and Mohit Iyyer. PEARL: Prompting large language models to plan and execute actions over long documents. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.29/.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F. Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, Hoang H. Tran, Fuqiang Li, Ren Ma, Mingzhang Zheng, Bill Qian, Yanjun Shao, Niklas Muennighoff, Yizhe Zhang, Binyuan Hui, Junyang Lin, Robert Brennan, Hao Peng, Heng Ji, and Graham Neubig. Openhands: An open platform for AI software developers as generalist agents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=OJd3ayDDoF.

Tianbao Xie, Fan Zhou, Zhoujun Cheng, Peng Shi, Luoxuan Weng, Yitao Liu, Toh Jing Hua, Junning Zhao, Qian Liu, Che Liu, et al. Openagents: An open platform for language agents in the wild. *arXiv preprint arXiv:2310.10634*, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37:50528–50652, 2024.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*, 2024.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Arik. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37:132208–132237, 2024.

Jun Zhao, Can Zu, Xu Hao, Yi Lu, Wei He, Yiwen Ding, Tao Gui, Qi Zhang, and Xuanjing Huang. LONGAGENT: Achieving question answering for 128k-token-long documents through multi-agent collaboration. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.912/.

Lee B Ziffer. The robots are coming: Ai large language models and the legal profession. In *American Bar Association*, 2023.

A Large Language Model Usage in Paper Writing

We use LLMs solely for language polishing purposes: grammar correction and paraphrasing to improve clarity and readability. We do not use LLMs to generate new content. All semantic content and scientific contributions originate entirely from the authors.

B Checklist Definitions

The following are the definitions of the 26 checklist items used in our work, which are adapted from ExpertLongBench (Ruan et al., 2025). We group them into 9 groups.

A. Basic Case Information

1. **Filing Date**: The date when the lawsuit was first initiated with the court.
2. **Parties**: Description of each plaintiff and defendant involved, including relevant positions or offices held. Use specific terms (e.g., “The city”, “The parents”) rather than generic terms (e.g., “The defendant”, “The plaintiffs”).
3. **Class Action or Individual Plaintiffs**: Whether the case involves class action plaintiffs or individual plaintiffs with descriptions.
4. **Type of Counsel**: The type(s) of counsel representing each side. Use brief category labels (e.g., private counsel, public interest nonprofit, government counsel, pro se) and include specific organizations (if applicable) in parentheses (e.g., Public interest nonprofit (ACLU)).

B. Legal Foundation

5. **Cause of Action**: The legal vehicle(s) used to bring the claims (the “how” of suing), such as statutes that create a private/enforcement right of action (e.g., 42 U.S.C. § 1983, Title II ADA, FTCA) or judge-made vehicles (e.g., Bivens).
6. **Statutory/Constitutional Basis**: The substantive rights and sources of law allegedly violated (the ’what’ was violated), such as specific constitutional provisions/clauses (e.g., Fourteenth Amendment—Equal Protection, First Amendment—Freedom of Association, Eighth Amendment) and statutory rights (e.g., ADA Title II, Rehab Act § 504).
7. **Remedy Sought**: What each party asks the court to grant, not what the court ordered or what the parties settled. Include both sides if the defendant seeks relief.

C. Judge Information

8. **Judge Name**: The first and last name of the judge(s) involved in the case. Do not include Supreme Court Justices.

D. Related Cases

9. **Consolidated Cases**: Cases that were combined with this case for joint proceedings.
10. **Related Cases**: Other cases referenced or connected to this case, listed by case code number.

E. Filings and Proceedings

11. **Important Filings**: Significant motions filed, including temporary restraining orders, preliminary injunctions, motions to dismiss, and motions for summary judgment.
12. **Court Rulings**: Judicial decisions on important filings such as motions to dismiss, summary judgment, preliminary injunctions, class certification, and attorneys’ fees (excluding amended complaints and statements of interest).
13. **Reported Opinions**: Citations of reported opinions using shortened Bluebook format (e.g., “2020 WL 4218003”), without case name, court, or date unless from a different case.
14. **Trials**: Information about trial proceedings including scheduling, outcomes, and related motions or rulings.
15. **Appeals**: Whether appeals were filed, which parties appealed, to which court, and the outcomes.

F. Decrees

16. **Significant Terms**: The substantive obligations ordered by the court. This includes consent decrees and stipulated judgments/injunctions because they are entered as court orders.

17. **Decree Dates**: All decree-related dates such as entry date, modification/amendment dates (of the order), suspension/stay dates, partial termination dates, and full termination/vacatur dates. Decrees include injunctions, consent decrees, or stipulated judgments/injunctions.
18. **Duration**: The duration of all decrees obligations (each as a separate entry). A 'decree' is any formal order or judgment issued by a court such as an injunction, consent decree, or stipulated judgment/injunction, as opposed to a negotiated agreement between parties.

G. Settlements

19. **Settlement Terms**: The substantive obligations the parties agree to in a settlement that is not entered as a court order. A settlement may be court-approved or enforced, but as long as it is not entered as an order, it is a settlement.
20. **Settlement Date**: All settlement-related dates (each as a separate entry) such as execution/signing date(s), court approval date (if approved but not entered as an order), amendment dates, enforcement/retention dates without incorporation (e.g., court retains jurisdiction over the settlement but does not enter it as an order), and termination/expiration of the settlement agreement (if contractual).
21. **Duration**: The duration of all settlements obligations (each as a separate entry). A 'settlement' is any negotiated agreement between parties that resolves a dispute, as opposed to a formal order or judgment issued by a court.
22. **Court Enforcement**: Whether the settlement (not entered as an order/judgment) is court-enforced. Answer Yes if the court explicitly retains jurisdiction to enforce the settlement without incorporating it into an order/judgment (e.g., Kokkonen retention). Answer No if it's a private agreement with no retained jurisdiction.
23. **Enforcement Disputes**: The disputes about enforcing a settlement (a negotiated agreement not entered as a court order)—e.g., motions to enforce/contempt or requests invoking retained jurisdiction—each as a separate value with date, movant, issue, and outcome (or pending).

H. Monitoring

24. **Monitor Name**: Name of any court-appointed monitor or special master.
25. **Monitor Reports**: Monitor's findings regarding defendant compliance with court orders, including which terms are being met.

I. Context

26. **Factual Basis**: The underlying facts and evidence supporting the legal claims, including: (i) details of relevant events (what, when, where, who), (ii) supporting evidence (physical, documentary, testimonial), and (iii) background context.

C Writing Style Similarity Evaluation Details

The following are the definitions of the five aspects used in our writing style similarity evaluation. Each aspect is rated on a 1–5 Likert scale, where 5 indicates identical and 1 indicates completely different.

1. **Readability & Jargon Level**
   Compare the reading level and the balance of legal jargon vs. plain language. Consider terminology density and accessibility to non-legal readers.
   - 5 Nearly identical reading level and jargon density; same balance of technical/plain language throughout.
   - 4 Very similar complexity with minor differences in terminology or occasional variance in technical language.
   - 3 Moderate differences in accessibility; one is noticeably more technical in places but overall similar.
   - 2 Significantly different complexity; one is consistently more technical or more accessible.
   - 1 Completely different target audiences (e.g., one for legal professionals, the other for the general public).
2. **Narrative Order**
   Compare whether events are presented in the same sequence (chronological vs. thematic) and the ordering of key facts and arguments.

5 Identical sequence of information; same events, facts, and arguments in the same order.
4 Same overall flow with 1–2 elements reordered; core structure preserved.
3 Similar general structure but several sections reordered; recognizable yet rearranged.
2 Different organizational approaches with some overlap (mix of chronological and thematic).
1 Completely different information architecture (e.g., one chronological, the other organized by issues).

3. **Sentence Structure & Voice**
Compare sentence variety, active vs. passive voice, and tense consistency.

5 Nearly identical sentence patterns, voice usage, and tense choices throughout.
4 Very similar style with occasional differences in sentence complexity or voice.
3 Moderate variation; one favors longer/shorter sentences or more active/passive constructions.
2 Noticeably different styles; consistent differences in sentence variety and voice preferences.
1 Completely different approaches (e.g., one varied and active; the other uniform and passive).

4. **Formatting & Layout**
Compare use of headings, bullet/numbered lists, paragraphing, and other structural cues.

5 Identical formatting choices; same use of headings, lists, and paragraph breaks.
4 Very similar structure with minor variations (e.g., one extra heading or different list style).
3 Similar approach but noticeable differences in execution (e.g., both use headings but at different levels/frequency).
2 Different formatting philosophies; one is much more structured than the other.
1 Completely different (e.g., one heavily formatted; the other continuous prose).

5. **Citation & Reference Style**
Compare presence, position, and formatting of case/statute citations or footnotes (inline vs. separate), citation density, and conventions.

5 Identical citation approach; same style, frequency, and positioning.
4 Very similar practices with minor formatting differences or occasional variation in placement.
3 Similar philosophy but different execution (e.g., both cite cases but differ in density/positioning).
2 Different approaches; one is substantially more reference-heavy or uses a different citation style.
1 Completely different or incomparable (e.g., one with extensive citations, the other with none).

D Annotation Details

Annotator Recruitment. We recruit four in-house annotators who are native English speakers and U.S.-based undergraduate students with basic familiarity with legal cases. All annotators are trained by the authors: we review the 26 checklist items together, ensure that everyone understands the legal terms involved (e.g., decree, settlement, ruling), and walk through example annotations. Because their task is to extract checklist items from case summaries that are written for lay readers rather than to provide legal judgments or read case documents, we do not require formal legal training once they clearly understand each checklist item and its definition.

Inter-Annotator Agreement. For checklist extraction, the ten longest summaries receive triple annotations. Agreement is measured as the average pairwise $S_{\text{checklist}}$ score across annotators, reaching 83.6 (using Gemma3 27B as the comparison model). For checklist comparison, single-value pairs achieve moderate agreement with Fleiss' $\kappa = 0.57$, while multi-value matching yields an average pairwise F1 of 0.82, indicating high consistency. For writing style similarity, Krippendorff's α (Krippendorff, 2011) across the five aspects averages 0.32. We also measure a "two-agree" metric: overall, at least two annotators agree with each other on the rating 94.4% of the time, and all three annotators choose different ratings only 5.6% of the time. This indicates that most instances of writing-style rating show clear majority agreement, and full disagreement is rare.

Annotation Interfaces. Figures 23, 24, and 25 display screenshots of our human annotation interfaces for checklist extraction, checklist comparison, and writing style similarity rating, respectively.

The collected data are used for the meta-evaluation of GAVEL-REF and for evaluating checklist extraction from case documents methods.

E FURTHER ANALYSIS

Figure 6 shows the average summary length of each LLM in each case-length bin, alongside the overall $S_{\text{GAVEL-REF}}$ score.

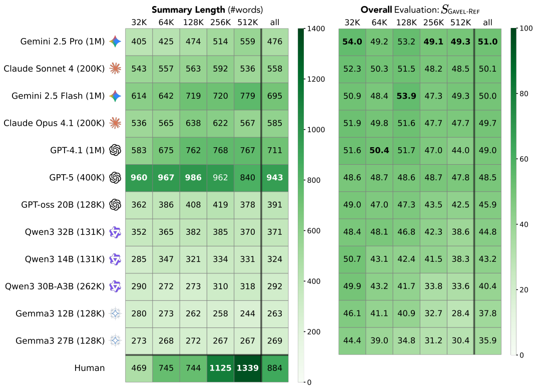


Figure 6: Summary length and overall evaluation for 12 LLMs. As case length increases, all models perform worse. For the cases in the 256K and 512K bins, LLM-generated summaries are much shorter than human summaries and fail to include as much information.

Compared to human summaries, LLMs only approach human length in the 32K–128K bins; for 256K and 512K cases, all models produce much shorter summaries than humans. In general, open-source models generate noticeably shorter summaries than proprietary models. Among all models, GPT5 is an outlier: it consistently produces very long summaries (often over 900 words) even for short cases (32K–128K), substantially longer than the human references. Figure 9 shows a typical example. GPT-5 often writes in a highly verbose, list-style format rather than a narrative, which we hypothesize is related to its “high” thinking mode. We also compute instance-level correlations between summary length and $S_{\text{GAVEL-REF}}$. Overall, we observe a moderate positive correlation (Pearson $r = 0.31$, Spearman $\rho = 0.36$, Kendall’s $\tau = 0.24$), but this is largely driven by weaker open-source models that both underperform and produce shorter summaries. When we separate proprietary and open-source models, the correlations become much smaller: within proprietary models, Pearson $r = -0.11$, Spearman $\rho = -0.13$, and Kendall’s $\tau = 0.09$; within open-source models, Pearson $r = 0.20$, Spearman $\rho = 0.20$, and Kendall’s $\tau = 0.14$. This suggests that, once we control for model family, summary length alone explains only a small fraction of the performance differences.

Figure 7 presents the item-level performance for the top 3 models in checklist evaluation—Gemini 2.5 Flash, Pro and Claude Sonnet 4—showing their top and bottom 5 checklist items plus consistently over- and under-specified items. All three models exhibit high similar performance patterns across items.

Figure 8 presents the checklist extraction performance $S_{\text{checklist}}$ versus total, input, output token usage for each method extracting checklist from case documents.

Figures 10, 11, and 12 present the checklist item-level performance for each of the 12 LLMs we evaluate.

Figures 13 to 22 show a randomly sampled case, comparing checklists extracted directly from case documents by GAVEL-AGENT with Qwen3 30B-A3B (26-agent configuration) against the human-annotated checklist extracted from the case summary.

F IMPLEMENTATION DETAILS

For all language models, we use a temperature of 0.7 and top-p of 1, except for GPT-5 (where temperature cannot be changed and is fixed at 1) and Qwen3, for which we use a temperature of 0.6 and top-p of 0.95, following the official recommendations. For Gemini 2.5 Flash and Pro, we set the thinking budget to -1 (allowing the model to decide). For GPT-5, we use “high” thinking effort. For Claude Sonnet 4 and Opus 4.1, we set the thinking budget to 10,000.

We use the following versions of the proprietary models: gpt-4.1-2025-04-14, gpt-5-2025-08-07, claude-sonnet-4-20250514, claude-opus-4-1-20250805, gemini-2.5-flash (June 2025), and gemini-

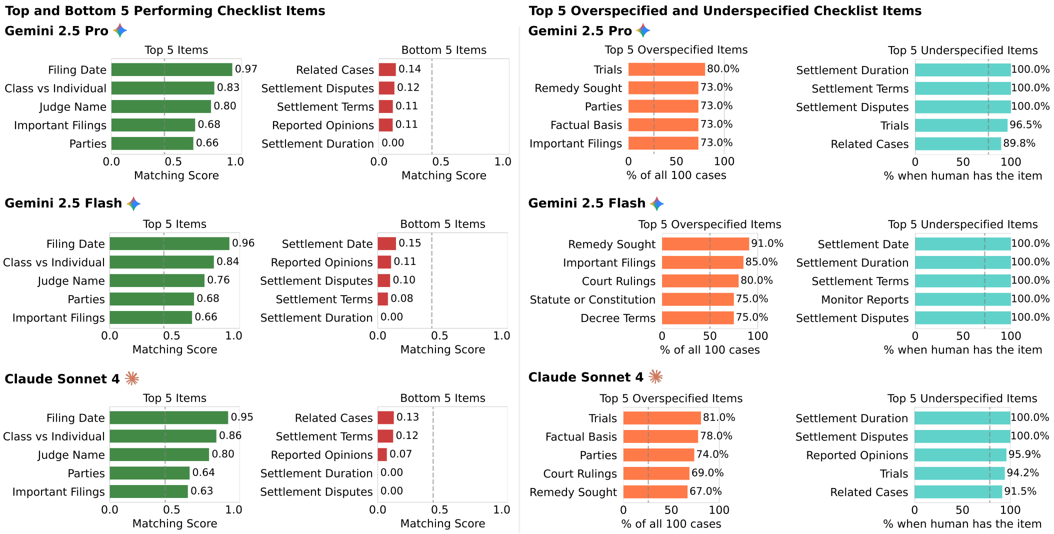


Figure 7: Performance breakdown for the top-3 models in checklist evaluation (Gemini 2.5 Pro, Gemini 2.5 Flash, and Claude Sonnet 4): top/bottom 5 checklist items by matching score and most frequently over/under-specified items. Overspecification measured as frequency across all 100 cases; underspecification as frequency among cases where human summary includes that item.

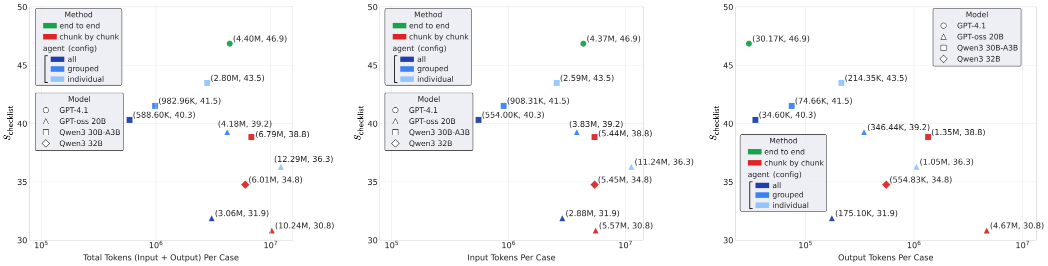


Figure 8: $S_{\text{checklist}}$ versus total token, input token, and output token usage for different methods extracting from case documents.

2.5-pro (June 2025). For open-source models, we use the instruction-tuned version of Gemma3 (Gemma3-it) and Qwen3-30B-A3B-Thinking-2507 for Qwen3 30B-A3B. Open-source models are run through vLLM Kwon et al. (2023) on 4 A40 GPUs. For all reasoning models such as Qwen3, we use the reasoning mode. Due to compute constraints, we could not run models larger than these, such as GPT-oss 120B. The total API costs is $1,800 USD.

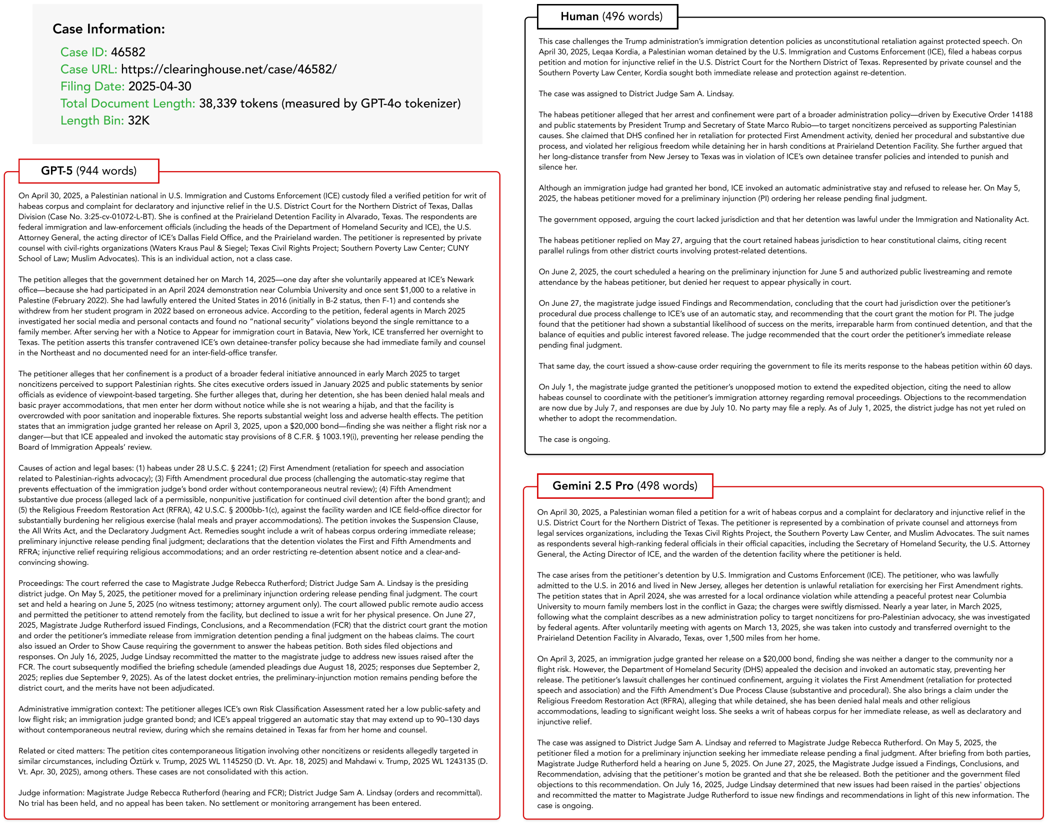
For GAVEL-AGENT, we implement tool calls using each model's native format: ChatML for Qwen3 and Harmony for GPT-oss.

G PROMPTS

The following lists the prompts used in our paper.

Prompts used in GAVEL-REF. Figure 26 shows the prompt for extracting checklist items from summaries. Figures 27 and 28 show the prompts for comparing single-value and multi-value checklist items, respectively. Figure 29 shows the prompt for extracting residual facts not covered by checklist items or their supporting text. Figure 30 shows the prompt for rating writing style similarity between two summaries across five aspects.

Prompt for summarization. Figure 31 shows the prompt for legal summarization.



Case Information:

Case ID: 46582
Case URL: https://clearinghouse.net/case/46582/
Filing Date: 2025-04-30
Total Document Length: 38,339 tokens (measured by GPT-4o tokenizer)
Length Bin: 32K

GPT-5 (944 words)

On April 30, 2025, a Palestinian national in U.S. Immigration and Customs Enforcement (ICE) custody filed a verified petition for writ of habeas corpus and complaint for declaratory and injunctive relief in the U.S. District Court for the Northern District of Texas, Dallas Division (Case No. 3:25-cv-01072-L-BT). She is confined at the Prairieland Detention Facility in Alvarado, Texas. The respondents are federal immigration and law-enforcement officials (including the heads of the Department of Homeland Security and ICE), the U.S. Attorney General, the acting director of ICE's Dallas Field Office, and the Prairieland warden. The petitioner is represented by private counsel with civil-rights organizations (Waters Kraus Paul & Siegel; Texas Civil Rights Project; Southern Poverty Law Center; CUNY School of Law; Muslim Advocates). This is an individual action, not a class case.

The petition alleges that the government detained her on March 14, 2025—one day after she voluntarily appeared at ICE's Newark office—because she had participated in an April 2024 demonstration near Columbia University and once sent $1,000 to a relative in Palestine (February 2022). She had lawfully entered the United States in 2016 (initially in B-2 status, then F-1) and contends she withdrew from her student program in 2022 based on erroneous advice. According to the petition, federal agents in March 2025 investigated her social media and personal contacts and found no "national security" violations beyond the single remittance to a family member. After serving her with a Notice to Appear for immigration court in Batavia, New York, ICE transferred her overnight to Texas. The petition asserts this transfer contravened ICE's own detainee-transfer policy because she had immediate family and counsel in the Northeast and no documented need for an inter-field-office transfer.

The petitioner alleges that her confinement is a product of a broader federal initiative announced in early March 2025 to target noncitizens perceived to support Palestinian rights. She cites executive orders issued in January 2025 and public statements by senior officials as evidence of viewpoint-based targeting. She further alleges that, during her detention, she has been denied halal meals and basic prayer accommodations, that men enter her dorm without notice while she is not wearing a hijab, and that the facility is overcrowded with poor sanitation and inoperable fixtures. She reports substantial weight loss and adverse health effects. The petition states that an immigration judge granted her release on April 3, 2025, upon a $20,000 bond—finding she was neither a flight risk nor a danger—but that ICE appealed and invoked the automatic stay provisions of 8 C.F.R. § 1003.19(i), preventing her release pending the Board of Immigration Appeals' review.

Causes of action and legal bases: (1) habeas under 28 U.S.C. § 2241; (2) First Amendment (retaliation for speech and association related to Palestinian-rights advocacy); (3) Fifth Amendment procedural due process (challenging the automatic-stay regime that prevents effectuation of the immigration judge's bond order without contemporaneous neutral review); (4) Fifth Amendment substantive due process (alleged lack of a permissible, nonpunitive justification for continued civil detention after the bond grant); and (5) the Religious Freedom Restoration Act (RFRA), 42 U.S.C. § 2000bb-1(c), against the facility warden and ICE field-office director for substantially burdening her religious exercise (halal meals and prayer accommodations). The petition invokes the Suspension Clause, the All Writs Act, and the Declaratory Judgment Act. Remedies sought include a writ of habeas corpus ordering immediate release; preliminary injunctive release pending final judgment; declarations that the detention violates the First and Fifth Amendments and RFRA; injunctive relief requiring religious accommodations; and an order restricting re-detention absent notice and a clear-and-convincing showing.

Proceedings: The court referred the case to Magistrate Judge Rebecca Rutherford; District Judge Sam A. Lindsay is the presiding district judge. On May 5, 2025, the petitioner moved for a preliminary injunction ordering release pending final judgment. The court set and held a hearing on June 5, 2025 (no witness testimony; attorney argument only). The court allowed public remote audio access and permitted the petitioner to attend remotely from the facility, but declined to issue a writ for her physical presence. On June 27, 2025, Magistrate Judge Rutherford issued Findings, Conclusions, and a Recommendation (FCR) that the district court grant the motion and order the petitioner's immediate release from immigration detention pending a final judgment on the habeas claims. The court also issued an Order to Show Cause requiring the government to answer the habeas petition. Both sides filed objections and responses. On July 16, 2025, Judge Lindsay recommitted the matter to the magistrate judge to address new issues raised after the FCR. The court subsequently modified the briefing schedule (amended pleadings due August 18, 2025; responses due September 2, 2025; replies due September 9, 2025). As of the latest docket entries, the preliminary-injunction motion remains pending before the district court, and the merits have not been adjudicated.

Administrative immigration context: The petitioner alleges ICE's own Risk Classification Assessment rated her a low public-safety and low flight risk; an immigration judge granted bond; and ICE's appeal triggered an automatic stay that may extend up to 90–130 days without contemporaneous neutral review, during which she remains detained in Texas far from her home and counsel.

Related or cited matters: The petition cites contemporaneous litigation involving other noncitizens or residents allegedly targeted in similar circumstances, including Öztürk v. Trump, 2025 WL 1145250 (D. Vt. Apr. 18, 2025) and Mahdawi v. Trump, 2025 WL 1243135 (D. Vt. Apr. 30, 2025), among others. These cases are not consolidated with this action.

Judge information: Magistrate Judge Rebecca Rutherford (hearing and FCR); District Judge Sam A. Lindsay (orders and recommittal). No trial has been held, and no appeal has been taken. No settlement or monitoring arrangement has been entered.

Human (496 words)

This case challenges the Trump administration's immigration detention policies as unconstitutional retaliation against protected speech. On April 30, 2025, Leqaa Kordia, a Palestinian woman detained by the U.S. Immigration and Customs Enforcement (ICE), filed a habeas corpus petition and motion for injunctive relief in the U.S. District Court for the Northern District of Texas. Represented by private counsel and the Southern Poverty Law Center, Kordia sought both immediate release and protection against re-detention.

The case was assigned to District Judge Sam A. Lindsay.

The habeas petitioner alleged that her arrest and confinement were part of a broader administration policy—driven by Executive Order 14188 and public statements by President Trump and Secretary of State Marco Rubio—to target noncitizens perceived as supporting Palestinian causes. She claimed that DHS confined her in retaliation for protected First Amendment activity, denied her procedural and substantive due process, and violated her religious freedom while detaining her in harsh conditions at Prairieland Detention Facility. She further argued that her long-distance transfer from New Jersey to Texas was in violation of ICE's own detainee transfer policies and intended to punish and silence her.

Although an immigration judge had granted her bond, ICE invoked an automatic administrative stay and refused to release her. On May 5, 2025, the habeas petitioner moved for a preliminary injunction (PI) ordering her release pending final judgment.

The government opposed, arguing the court lacked jurisdiction and that her detention was lawful under the Immigration and Nationality Act.

The habeas petitioner replied on May 27, arguing that the court retained habeas jurisdiction to hear constitutional claims, citing recent parallel rulings from other district courts involving protest-related detentions.

On June 2, 2025, the court scheduled a hearing on the preliminary injunction for June 5 and authorized public livestreaming and remote attendance by the habeas petitioner, but denied her request to appear physically in court.

On June 27, the magistrate judge issued Findings and Recommendation, concluding that the court had jurisdiction over the petitioner's procedural due process challenge to ICE's use of an automatic stay, and recommending that the court grant the motion for PI. The judge found that the petitioner had shown a substantial likelihood of success on the merits, irreparable harm from continued detention, and that the balance of equities and public interest favored release. The judge recommended that the court order the petitioner's immediate release pending final judgment.

That same day, the court issued a show-cause order requiring the government to file its merits response to the habeas petition within 60 days.

On July 1, the magistrate judge granted the petitioner's unopposed motion to extend the expedited objection, citing the need to allow habeas counsel to coordinate with the petitioner's immigration attorney regarding removal proceedings. Objections to the recommendation are now due by July 7, and responses are due by July 10. No party may file a reply. As of July 1, 2025, the district judge has not yet ruled on whether to adopt the recommendation.

The case is ongoing.

Gemini 2.5 Pro (498 words)

On April 30, 2025, a Palestinian woman filed a petition for a writ of habeas corpus and a complaint for declaratory and injunctive relief in the U.S. District Court for the Northern District of Texas. The petitioner is represented by a combination of private counsel and attorneys from legal services organizations, including the Texas Civil Rights Project, the Southern Poverty Law Center, and Muslim Advocates. The suit names as respondents several high-ranking federal officials in their official capacities, including the Secretary of Homeland Security, the U.S. Attorney General, the Acting Director of ICE, and the warden of the detention facility where the petitioner is held.

The case arises from the petitioner's detention by U.S. Immigration and Customs Enforcement (ICE). The petitioner, who was lawfully admitted to the U.S. in 2016 and lived in New Jersey, alleges her detention is unlawful retaliation for exercising her First Amendment rights. The petition states that in April 2024, she was arrested for a local ordinance violation while attending a peaceful protest near Columbia University to mourn family members lost in the conflict in Gaza; the charges were swiftly dismissed. Nearly a year later, in March 2025, following what the complaint describes as a new administration policy to target noncitizens for pro-Palestinian advocacy, she was investigated by federal agents. After voluntarily meeting with agents on March 13, 2025, she was taken into custody and transferred overnight to the Prairieland Detention Facility in Alvarado, Texas, over 1,500 miles from her home.

On April 3, 2025, an immigration judge granted her release on a $20,000 bond, finding she was neither a danger to the community nor a flight risk. However, the Department of Homeland Security (DHS) appealed the decision and invoked an automatic stay, preventing her release. The petitioner's lawsuit challenges her continued confinement, arguing it violates the First Amendment (retaliation for protected speech and association) and the Fifth Amendment's Due Process Clause (substantive and procedural). She also brings a claim under the Religious Freedom Restoration Act (RFRA), alleging that while detained, she has been denied halal meals and other religious accommodations, leading to significant weight loss. She seeks a writ of habeas corpus for her immediate release, as well as declaratory and injunctive relief.

The case was assigned to District Judge Sam A. Lindsay and referred to Magistrate Judge Rebecca Rutherford. On May 5, 2025, the petitioner filed a motion for a preliminary injunction seeking her immediate release pending a final judgment. After briefing from both parties, Magistrate Judge Rutherford held a hearing on June 5, 2025. On June 27, 2025, the Magistrate Judge issued a Findings, Conclusions, and Recommendation, advising that the petitioner's motion be granted and that she be released. Both the petitioner and the government filed objections to this recommendation. On July 16, 2025, Judge Lindsay determined that new issues had been raised in the parties' objections and recommitted the matter to Magistrate Judge Rutherford to issue new findings and recommendations in light of this new information. The case is ongoing.

Figure 9: Example summaries from GPT-5, Gemini 2.5 Pro, and a human reference for a case in the 32K bin. This illustrates why GPT-5 produces very long summaries (as seen in Figure 6) even for short cases.

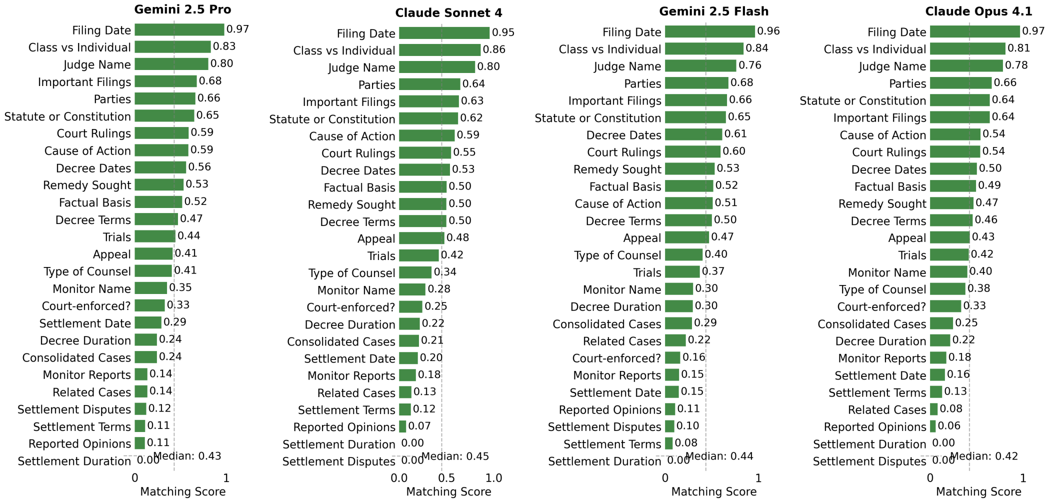


Figure 10: Checklist item-level performance for each LLM in the checklist evaluation. The metric is the matching score m_i. This figure shows results for Gemini 2.5 Pro, Claude Sonnet 4, Gemini 2.5 Flash, and Claude Opus 4.1.

Prompts for checklist extraction from case documents. Figures 32 and 33 present the prompts for the end-to-end method. Figure 34 presents the prompt for the chunk-by-chunk method. Figures 35, 36, and 37 present the system prompts used in GAVEL-AGENT.

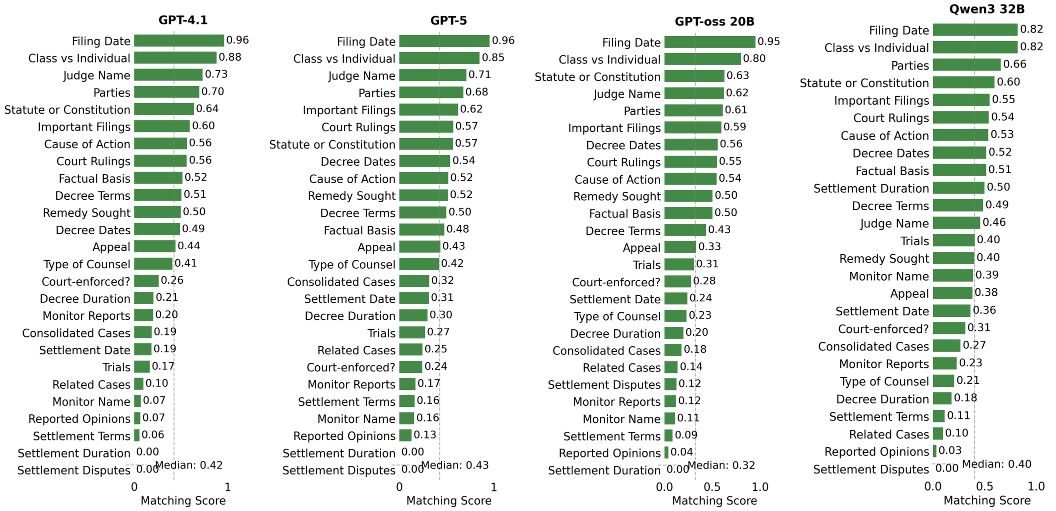


Figure 11: Checklist item-level performance for each LLM in the checklist evaluation. The metric is the matching score m_i. This figure shows results for GPT-4.1, GPT-5, GPT-oss 20B, Qwen3 32B.

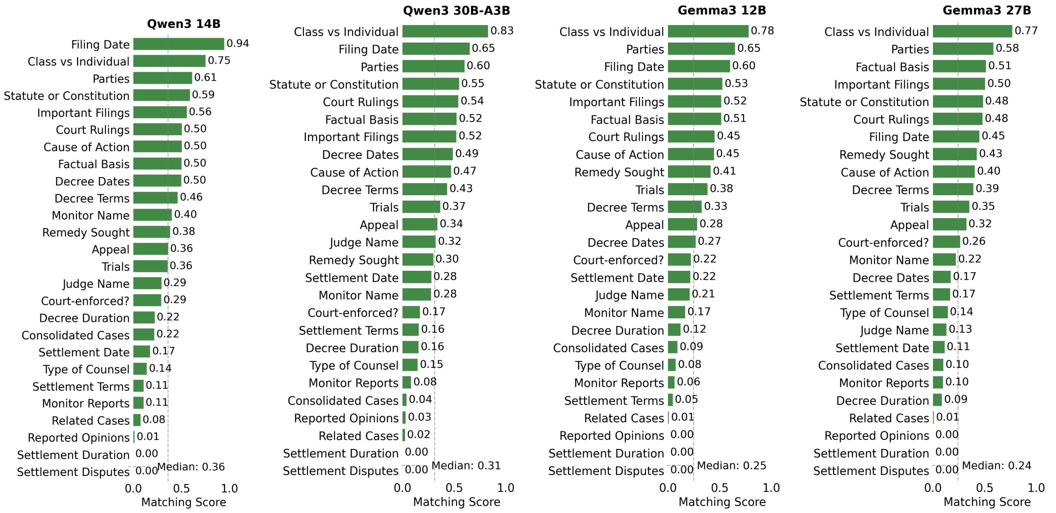
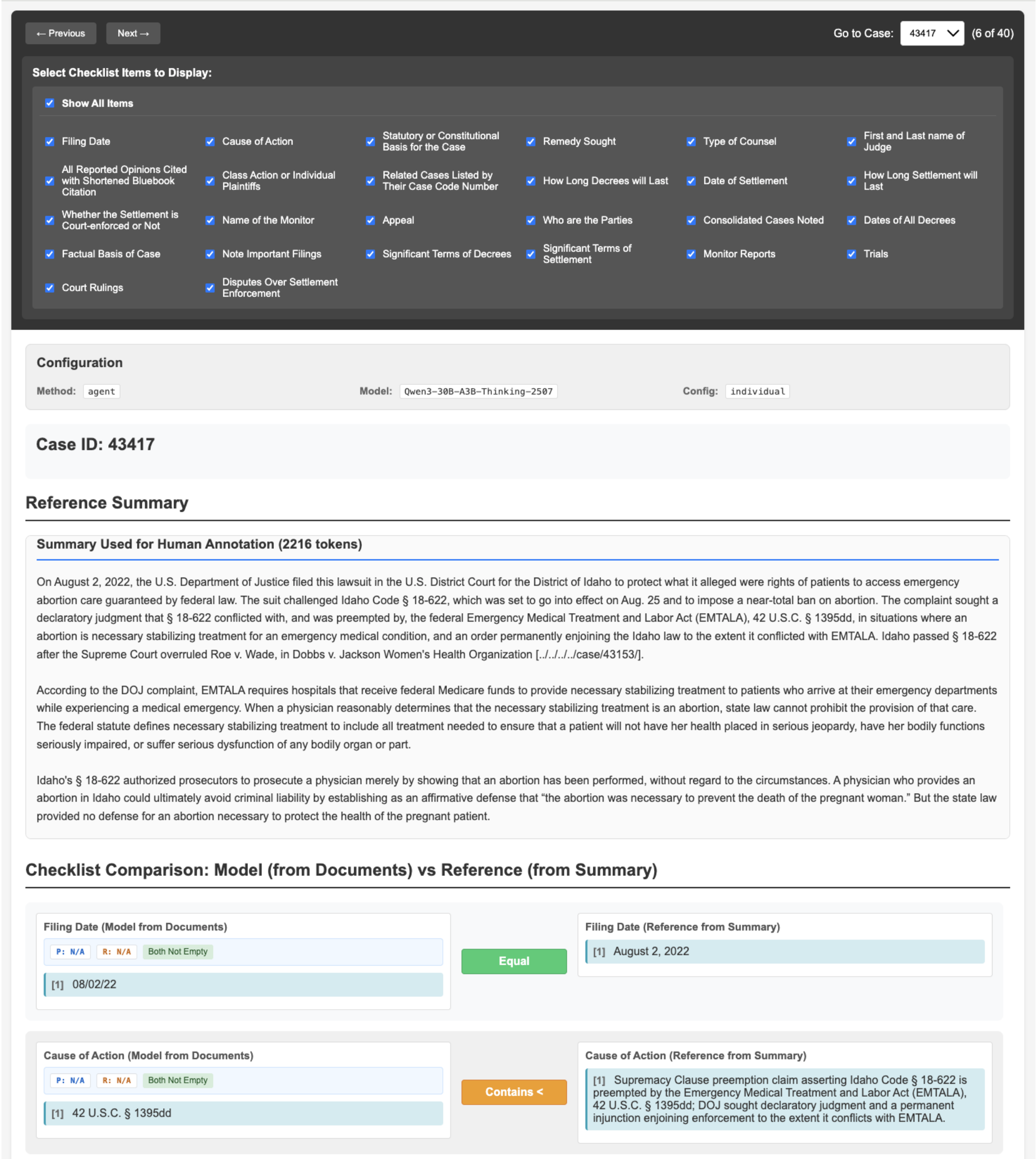


Figure 12: Checklist item-level performance for each LLM in the checklist evaluation. The metric is the matching score m_i. This figure shows results for Qwen3 14B, Qwen3 30B-A3B, Gemma3 12B and Gemma3 27B.



Select Checklist Items to Display:

☑ **Show All Items**

☑ Filing Date
☑ Cause of Action
☑ Statutory or Constitutional Basis for the Case
☑ Remedy Sought
☑ Type of Counsel
☑ First and Last name of Judge
☑ All Reported Opinions Cited with Shortened Bluebook Citation
☑ Class Action or Individual Plaintiffs
☑ Related Cases Listed by Their Case Code Number
☑ How Long Decrees will Last
☑ Date of Settlement
☑ How Long Settlement will Last
☑ Whether the Settlement is Court-enforced or Not
☑ Name of the Monitor
☑ Appeal
☑ Who are the Parties
☑ Consolidated Cases Noted
☑ Dates of All Decrees
☑ Factual Basis of Case
☑ Note Important Filings
☑ Significant Terms of Decrees
☑ Significant Terms of Settlement
☑ Monitor Reports
☑ Trials
☑ Court Rulings
☑ Disputes Over Settlement Enforcement

Configuration

Method: `agent` **Model:** `Qwen3-30B-A3B-Thinking-2507` **Config:** `individual`

Case ID: 43417

Reference Summary

Summary Used for Human Annotation (2216 tokens)

On August 2, 2022, the U.S. Department of Justice filed this lawsuit in the U.S. District Court for the District of Idaho to protect what it alleged were rights of patients to access emergency abortion care guaranteed by federal law. The suit challenged Idaho Code § 18-622, which was set to go into effect on Aug. 25 and to impose a near-total ban on abortion. The complaint sought a declaratory judgment that § 18-622 conflicted with, and was preempted by, the federal Emergency Medical Treatment and Labor Act (EMTALA), 42 U.S.C. § 1395dd, in situations where an abortion is necessary stabilizing treatment for an emergency medical condition, and an order permanently enjoining the Idaho law to the extent it conflicted with EMTALA. Idaho passed § 18-622 after the Supreme Court overruled Roe v. Wade, in Dobbs v. Jackson Women's Health Organization [../../../../case/43153/].

According to the DOJ complaint, EMTALA requires hospitals that receive federal Medicare funds to provide necessary stabilizing treatment to patients who arrive at their emergency departments while experiencing a medical emergency. When a physician reasonably determines that the necessary stabilizing treatment is an abortion, state law cannot prohibit the provision of that care. The federal statute defines necessary stabilizing treatment to include all treatment needed to ensure that a patient will not have her health placed in serious jeopardy, have her bodily functions seriously impaired, or suffer serious dysfunction of any bodily organ or part.

Idaho's § 18-622 authorized prosecutors to prosecute a physician merely by showing that an abortion has been performed, without regard to the circumstances. A physician who provides an abortion in Idaho could ultimately avoid criminal liability by establishing as an affirmative defense that "the abortion was necessary to prevent the death of the pregnant woman." But the state law provided no defense for an abortion necessary to protect the health of the pregnant patient.

Checklist Comparison: Model (from Documents) vs Reference (from Summary)

Filing Date (Model from Documents)

`P: N/A` `R: N/A` Both Not Empty

[1] 08/02/22

Equal

Filing Date (Reference from Summary)

[1] August 2, 2022

Cause of Action (Model from Documents)

`P: N/A` `R: N/A` Both Not Empty

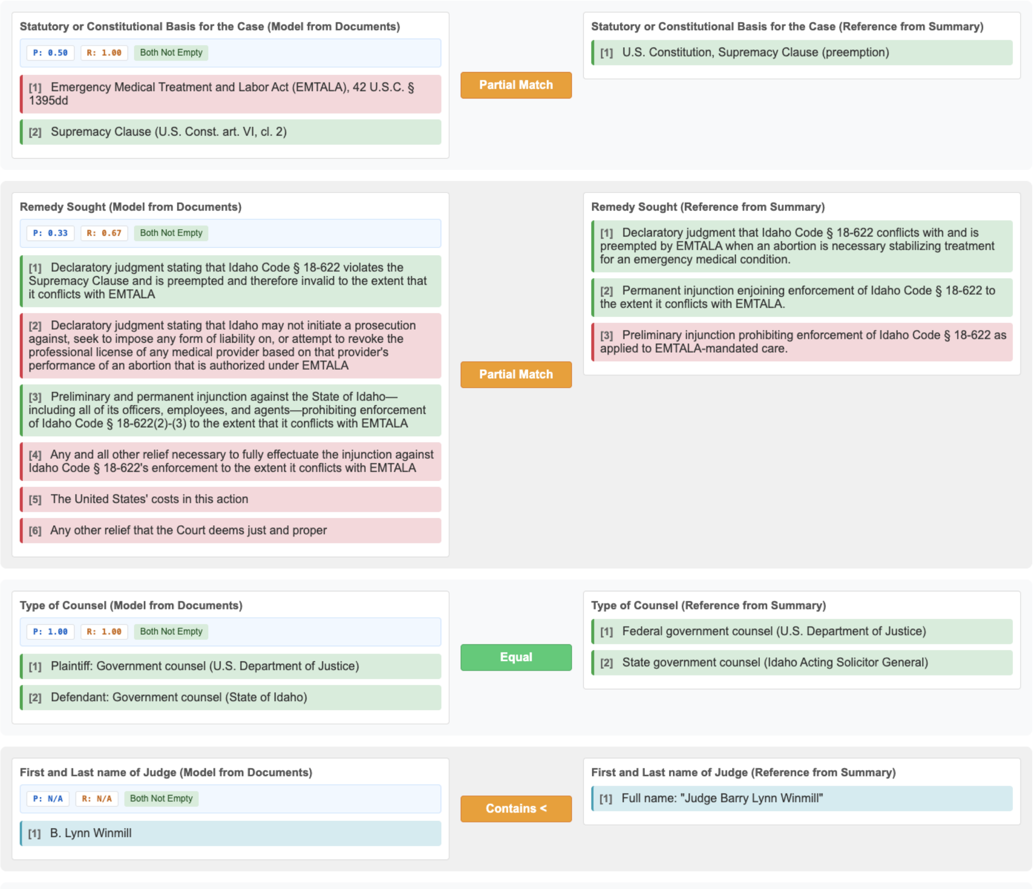
[1] 42 U.S.C. § 1395dd

Contains <

Cause of Action (Reference from Summary)

[1] Supremacy Clause preemption claim asserting Idaho Code § 18-622 is preempted by the Emergency Medical Treatment and Labor Act (EMTALA), 42 U.S.C. § 1395dd; DOJ sought declaratory judgment and a permanent injunction enjoining enforcement to the extent it conflicts with EMTALA.

Figure 13: Screenshot of a visualization for one case, comparing checklists extracted directly from case documents by GAVEL-AGENT with Qwen3 30B-A3B (26 individual agents configuration) against the human-annotated checklist extracted from the case summary (figure 1 of 10).



Statutory or Constitutional Basis for the Case (Model from Documents)

P: 0.50 R: 1.00 Both Not Empty

[1] Emergency Medical Treatment and Labor Act (EMTALA), 42 U.S.C. § 1395dd

[2] Supremacy Clause (U.S. Const. art. VI, cl. 2)

Partial Match

Statutory or Constitutional Basis for the Case (Reference from Summary)

[1] U.S. Constitution, Supremacy Clause (preemption)

Remedy Sought (Model from Documents)

P: 0.33 R: 0.67 Both Not Empty

[1] Declaratory judgment stating that Idaho Code § 18-622 violates the Supremacy Clause and is preempted and therefore invalid to the extent that it conflicts with EMTALA

[2] Declaratory judgment stating that Idaho may not initiate a prosecution against, seek to impose any form of liability on, or attempt to revoke the professional license of any medical provider based on that provider's performance of an abortion that is authorized under EMTALA

[3] Preliminary and permanent injunction against the State of Idaho—including all of its officers, employees, and agents—prohibiting enforcement of Idaho Code § 18-622(2)-(3) to the extent that it conflicts with EMTALA

[4] Any and all other relief necessary to fully effectuate the injunction against Idaho Code § 18-622's enforcement to the extent it conflicts with EMTALA

[5] The United States' costs in this action

[6] Any other relief that the Court deems just and proper

Partial Match

Remedy Sought (Reference from Summary)

[1] Declaratory judgment that Idaho Code § 18-622 conflicts with and is preempted by EMTALA when an abortion is necessary stabilizing treatment for an emergency medical condition.

[2] Permanent injunction enjoining enforcement of Idaho Code § 18-622 to the extent it conflicts with EMTALA.

[3] Preliminary injunction prohibiting enforcement of Idaho Code § 18-622 as applied to EMTALA-mandated care.

Type of Counsel (Model from Documents)

P: 1.00 R: 1.00 Both Not Empty

[1] Plaintiff: Government counsel (U.S. Department of Justice)

[2] Defendant: Government counsel (State of Idaho)

Equal

Type of Counsel (Reference from Summary)

[1] Federal government counsel (U.S. Department of Justice)

[2] State government counsel (Idaho Acting Solicitor General)

First and Last name of Judge (Model from Documents)

P: N/A R: N/A Both Not Empty

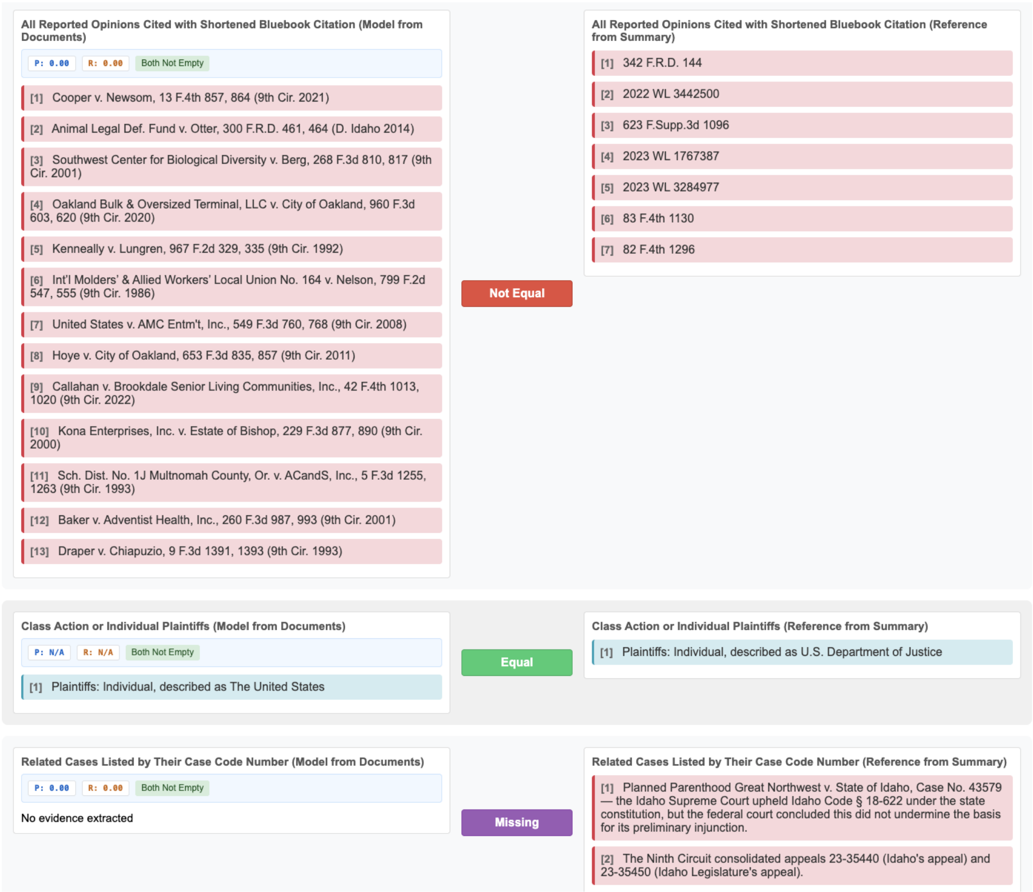
[1] B. Lynn Winmill

Contains <

First and Last name of Judge (Reference from Summary)

[1] Full name: "Judge Barry Lynn Winmill"

Figure 14: Screenshot of a visualization for one case, comparing checklists extracted directly from case documents by GAVEL-AGENT with Qwen3 30B-A3B (26 individual agents configuration) against the human-annotated checklist extracted from the case summary (figure 2 of 10).



All Reported Opinions Cited with Shortened Bluebook Citation (Model from Documents)

P: 0.00 R: 0.00 Both Not Empty

[1] Cooper v. Newsom, 13 F.4th 857, 864 (9th Cir. 2021)

[2] Animal Legal Def. Fund v. Otter, 300 F.R.D. 461, 464 (D. Idaho 2014)

[3] Southwest Center for Biological Diversity v. Berg, 268 F.3d 810, 817 (9th Cir. 2001)

[4] Oakland Bulk & Oversized Terminal, LLC v. City of Oakland, 960 F.3d 603, 620 (9th Cir. 2020)

[5] Kenneally v. Lungren, 967 F.2d 329, 335 (9th Cir. 1992)

[6] Int'l Molders' & Allied Workers' Local Union No. 164 v. Nelson, 799 F.2d 547, 555 (9th Cir. 1986)

[7] United States v. AMC Entm't, Inc., 549 F.3d 760, 768 (9th Cir. 2008)

[8] Hoye v. City of Oakland, 653 F.3d 835, 857 (9th Cir. 2011)

[9] Callahan v. Brookdale Senior Living Communities, Inc., 42 F.4th 1013, 1020 (9th Cir. 2022)

[10] Kona Enterprises, Inc. v. Estate of Bishop, 229 F.3d 877, 890 (9th Cir. 2000)

[11] Sch. Dist. No. 1J Multnomah County, Or. v. ACandS, Inc., 5 F.3d 1255, 1263 (9th Cir. 1993)

[12] Baker v. Adventist Health, Inc., 260 F.3d 987, 993 (9th Cir. 2001)

[13] Draper v. Chiapuzio, 9 F.3d 1391, 1393 (9th Cir. 1993)

Not Equal

All Reported Opinions Cited with Shortened Bluebook Citation (Reference from Summary)

[1] 342 F.R.D. 144

[2] 2022 WL 3442500

[3] 623 F.Supp.3d 1096

[4] 2023 WL 1767387

[5] 2023 WL 3284977

[6] 83 F.4th 1130

[7] 82 F.4th 1296

Class Action or Individual Plaintiffs (Model from Documents)

P: N/A R: N/A Both Not Empty

[1] Plaintiffs: Individual, described as The United States

Equal

Class Action or Individual Plaintiffs (Reference from Summary)

[1] Plaintiffs: Individual, described as U.S. Department of Justice

Related Cases Listed by Their Case Code Number (Model from Documents)

P: 0.00 R: 0.00 Both Not Empty

No evidence extracted

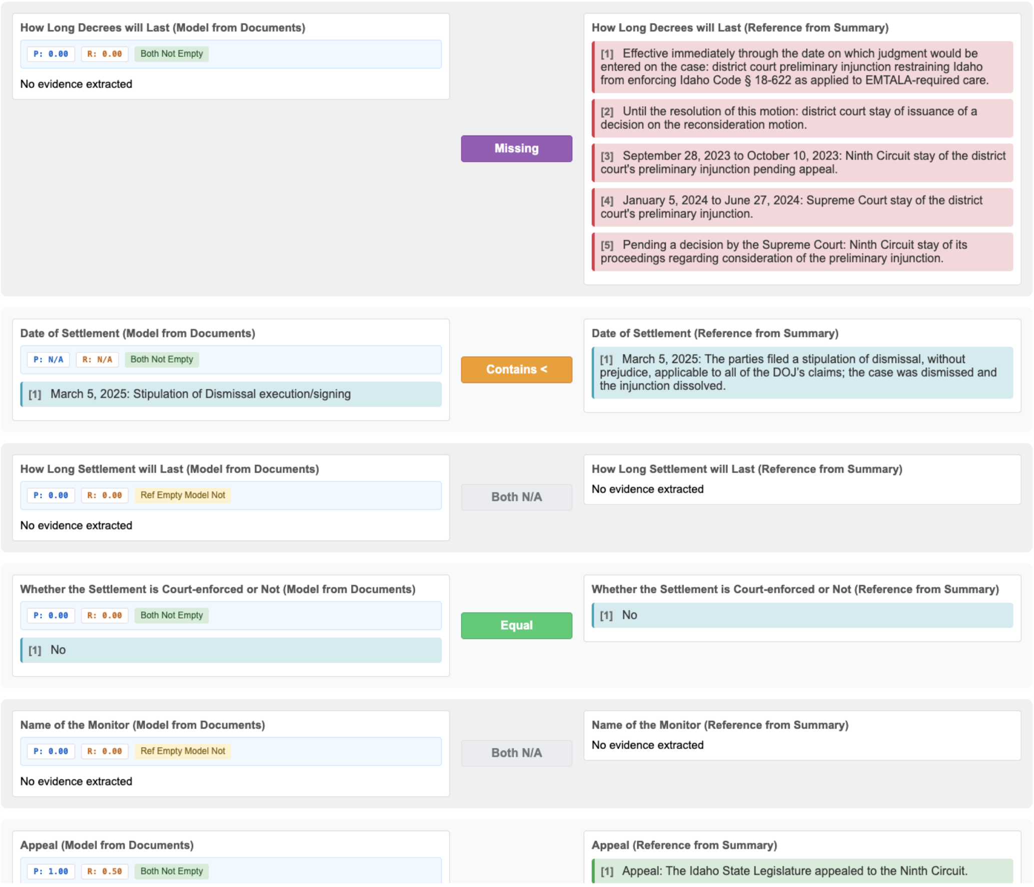
Missing

Related Cases Listed by Their Case Code Number (Reference from Summary)

[1] Planned Parenthood Great Northwest v. State of Idaho, Case No. 43579 — the Idaho Supreme Court upheld Idaho Code § 18-622 under the state constitution, but the federal court concluded this did not undermine the basis for its preliminary injunction.

[2] The Ninth Circuit consolidated appeals 23-35440 (Idaho's appeal) and 23-35450 (Idaho Legislature's appeal).

Figure 15: Screenshot of a visualization for one case, comparing checklists extracted directly from case documents by GAVEL-AGENT with Qwen3 30B-A3B (26 individual agents configuration) against the human-annotated checklist extracted from the case summary (figure 3 of 10).



How Long Decrees will Last (Model from Documents)

P: 0.00 R: 0.00 Both Not Empty

No evidence extracted

How Long Decrees will Last (Reference from Summary)

[1] Effective immediately through the date on which judgment would be entered on the case: district court preliminary injunction restraining Idaho from enforcing Idaho Code § 18-622 as applied to EMTALA-required care.

[2] Until the resolution of this motion: district court stay of issuance of a decision on the reconsideration motion.

[3] September 28, 2023 to October 10, 2023: Ninth Circuit stay of the district court's preliminary injunction pending appeal.

[4] January 5, 2024 to June 27, 2024: Supreme Court stay of the district court's preliminary injunction.

[5] Pending a decision by the Supreme Court: Ninth Circuit stay of its proceedings regarding consideration of the preliminary injunction.

Date of Settlement (Model from Documents)

P: N/A R: N/A Both Not Empty

[1] March 5, 2025: Stipulation of Dismissal execution/signing

Date of Settlement (Reference from Summary)

[1] March 5, 2025: The parties filed a stipulation of dismissal, without prejudice, applicable to all of the DOJ's claims; the case was dismissed and the injunction dissolved.

How Long Settlement will Last (Model from Documents)

P: 0.00 R: 0.00 Ref Empty Model Not

No evidence extracted

How Long Settlement will Last (Reference from Summary)

No evidence extracted

Whether the Settlement is Court-enforced or Not (Model from Documents)

P: 0.00 R: 0.00 Both Not Empty

[1] No

Whether the Settlement is Court-enforced or Not (Reference from Summary)

[1] No

Name of the Monitor (Model from Documents)

P: 0.00 R: 0.00 Ref Empty Model Not

No evidence extracted

Name of the Monitor (Reference from Summary)

No evidence extracted

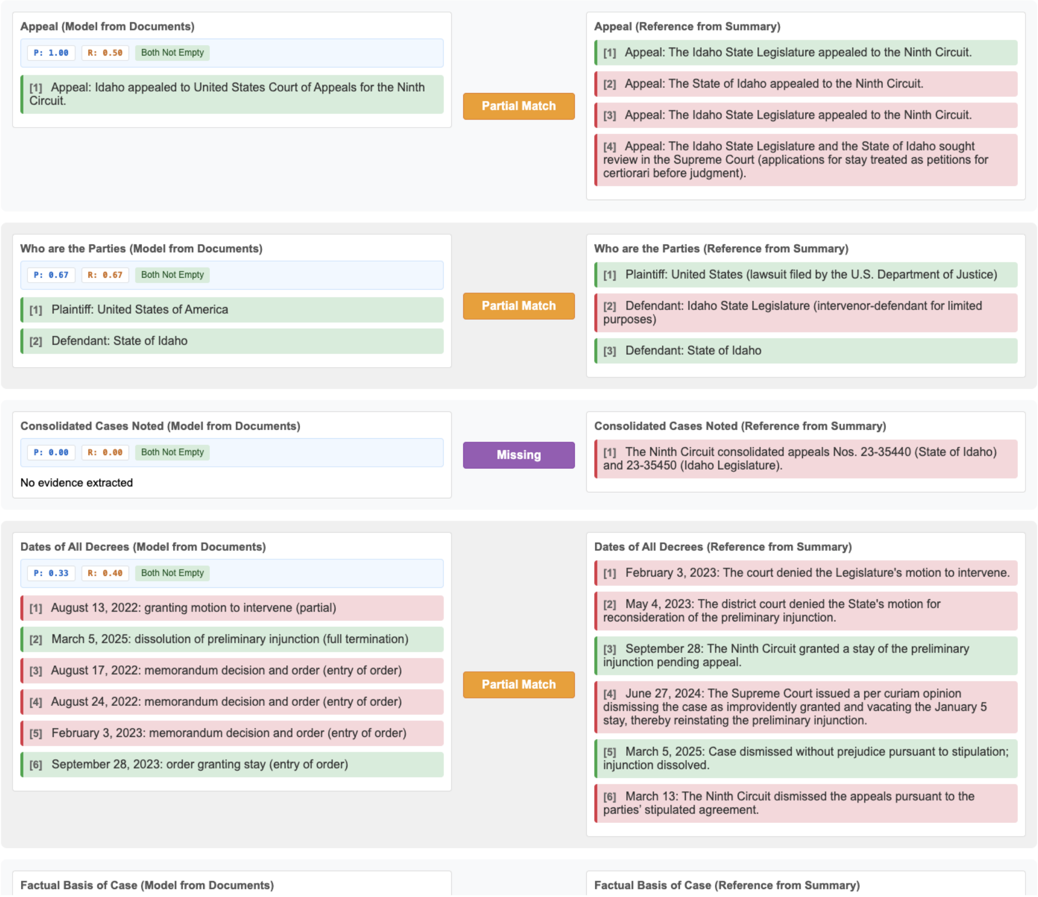
Appeal (Model from Documents)

P: 1.00 R: 0.50 Both Not Empty

Appeal (Reference from Summary)

[1] Appeal: The Idaho State Legislature appealed to the Ninth Circuit.

Figure 16: Screenshot of a visualization for one case, comparing checklists extracted directly from case documents by GAVEL-AGENT with Qwen3 30B-A3B (26 individual agents configuration) against the human-annotated checklist extracted from the case summary (figure 4 of 10).



Appeal (Model from Documents)

P: 1.00 R: 0.50 Both Not Empty

[1] Appeal: Idaho appealed to United States Court of Appeals for the Ninth Circuit.

Partial Match

Appeal (Reference from Summary)

[1] Appeal: The Idaho State Legislature appealed to the Ninth Circuit.

[2] Appeal: The State of Idaho appealed to the Ninth Circuit.

[3] Appeal: The Idaho State Legislature appealed to the Ninth Circuit.

[4] Appeal: The Idaho State Legislature and the State of Idaho sought review in the Supreme Court (applications for stay treated as petitions for certiorari before judgment).

Who are the Parties (Model from Documents)

P: 0.67 R: 0.67 Both Not Empty

[1] Plaintiff: United States of America

[2] Defendant: State of Idaho

Partial Match

Who are the Parties (Reference from Summary)

[1] Plaintiff: United States (lawsuit filed by the U.S. Department of Justice)

[2] Defendant: Idaho State Legislature (intervenor-defendant for limited purposes)

[3] Defendant: State of Idaho

Consolidated Cases Noted (Model from Documents)

P: 0.00 R: 0.00 Both Not Empty

No evidence extracted

Missing

Consolidated Cases Noted (Reference from Summary)

[1] The Ninth Circuit consolidated appeals Nos. 23-35440 (State of Idaho) and 23-35450 (Idaho Legislature).

Dates of All Decrees (Model from Documents)

P: 0.33 R: 0.40 Both Not Empty

[1] August 13, 2022: granting motion to intervene (partial)

[2] March 5, 2025: dissolution of preliminary injunction (full termination)

[3] August 17, 2022: memorandum decision and order (entry of order)

[4] August 24, 2022: memorandum decision and order (entry of order)

[5] February 3, 2023: memorandum decision and order (entry of order)

[6] September 28, 2023: order granting stay (entry of order)

Partial Match

Dates of All Decrees (Reference from Summary)

[1] February 3, 2023: The court denied the Legislature's motion to intervene.

[2] May 4, 2023: The district court denied the State's motion for reconsideration of the preliminary injunction.

[3] September 28: The Ninth Circuit granted a stay of the preliminary injunction pending appeal.

[4] June 27, 2024: The Supreme Court issued a per curiam opinion dismissing the case as improvidently granted and vacating the January 5 stay, thereby reinstating the preliminary injunction.

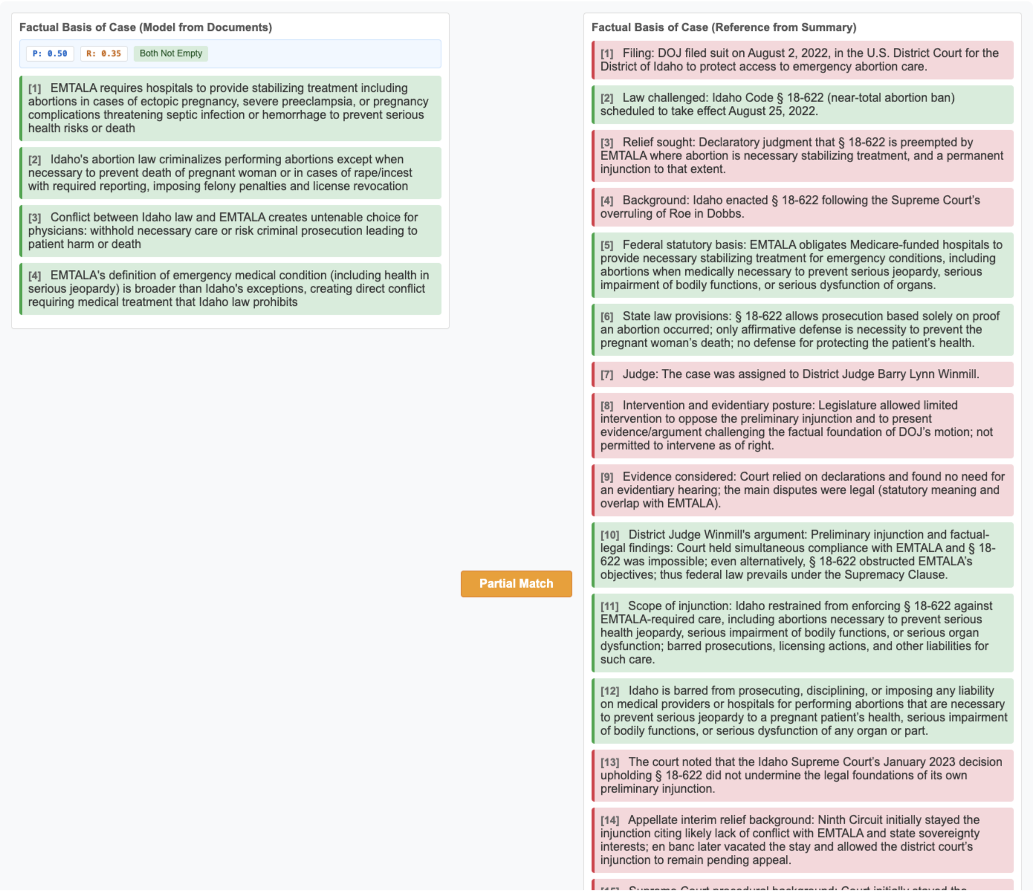
[5] March 5, 2025: Case dismissed without prejudice pursuant to stipulation; injunction dissolved.

[6] March 13: The Ninth Circuit dismissed the appeals pursuant to the parties' stipulated agreement.

Factual Basis of Case (Model from Documents)

Factual Basis of Case (Reference from Summary)

Figure 17: Screenshot of a visualization for one case, comparing checklists extracted directly from case documents by GAVEL-AGENT with Qwen3 30B-A3B (26 individual agents configuration) against the human-annotated checklist extracted from the case summary (figure 5 of 10).



Factual Basis of Case (Model from Documents)

P: 0.50 R: 0.35 Both Not Empty

[1] EMTALA requires hospitals to provide stabilizing treatment including abortions in cases of ectopic pregnancy, severe preeclampsia, or pregnancy complications threatening septic infection or hemorrhage to prevent serious health risks or death

[2] Idaho's abortion law criminalizes performing abortions except when necessary to prevent death of pregnant woman or in cases of rape/incest with required reporting, imposing felony penalties and license revocation

[3] Conflict between Idaho law and EMTALA creates untenable choice for physicians: withhold necessary care or risk criminal prosecution leading to patient harm or death

[4] EMTALA's definition of emergency medical condition (including health in serious jeopardy) is broader than Idaho's exceptions, creating direct conflict requiring medical treatment that Idaho law prohibits

Factual Basis of Case (Reference from Summary)

[1] Filing: DOJ filed suit on August 2, 2022, in the U.S. District Court for the District of Idaho to protect access to emergency abortion care.

[2] Law challenged: Idaho Code § 18-622 (near-total abortion ban) scheduled to take effect August 25, 2022.

[3] Relief sought: Declaratory judgment that § 18-622 is preempted by EMTALA where abortion is necessary stabilizing treatment, and a permanent injunction to that extent.

[4] Background: Idaho enacted § 18-622 following the Supreme Court's overruling of Roe in Dobbs.

[5] Federal statutory basis: EMTALA obligates Medicare-funded hospitals to provide necessary stabilizing treatment for emergency conditions, including abortions when medically necessary to prevent serious jeopardy, serious impairment of bodily functions, or serious dysfunction of organs.

[6] State law provisions: § 18-622 allows prosecution based solely on proof an abortion occurred; only affirmative defense is necessity to prevent the pregnant woman's death; no defense for protecting the patient's health.

[7] Judge: The case was assigned to District Judge Barry Lynn Winmill.

[8] Intervention and evidentiary posture: Legislature allowed limited intervention to oppose the preliminary injunction and to present evidence/argument challenging the factual foundation of DOJ's motion; not permitted to intervene as of right.

[9] Evidence considered: Court relied on declarations and found no need for an evidentiary hearing; the main disputes were legal (statutory meaning and overlap with EMTALA).

[10] District Judge Winmill's argument: Preliminary injunction and factual-legal findings: Court held simultaneous compliance with EMTALA and § 18-622 was impossible; even alternatively, § 18-622 obstructed EMTALA's objectives; thus federal law prevails under the Supremacy Clause.

[11] Scope of injunction: Idaho restrained from enforcing § 18-622 against EMTALA-required care, including abortions necessary to prevent serious health jeopardy, serious impairment of bodily functions, or serious organ dysfunction; barred prosecutions, licensing actions, and other liabilities for such care.

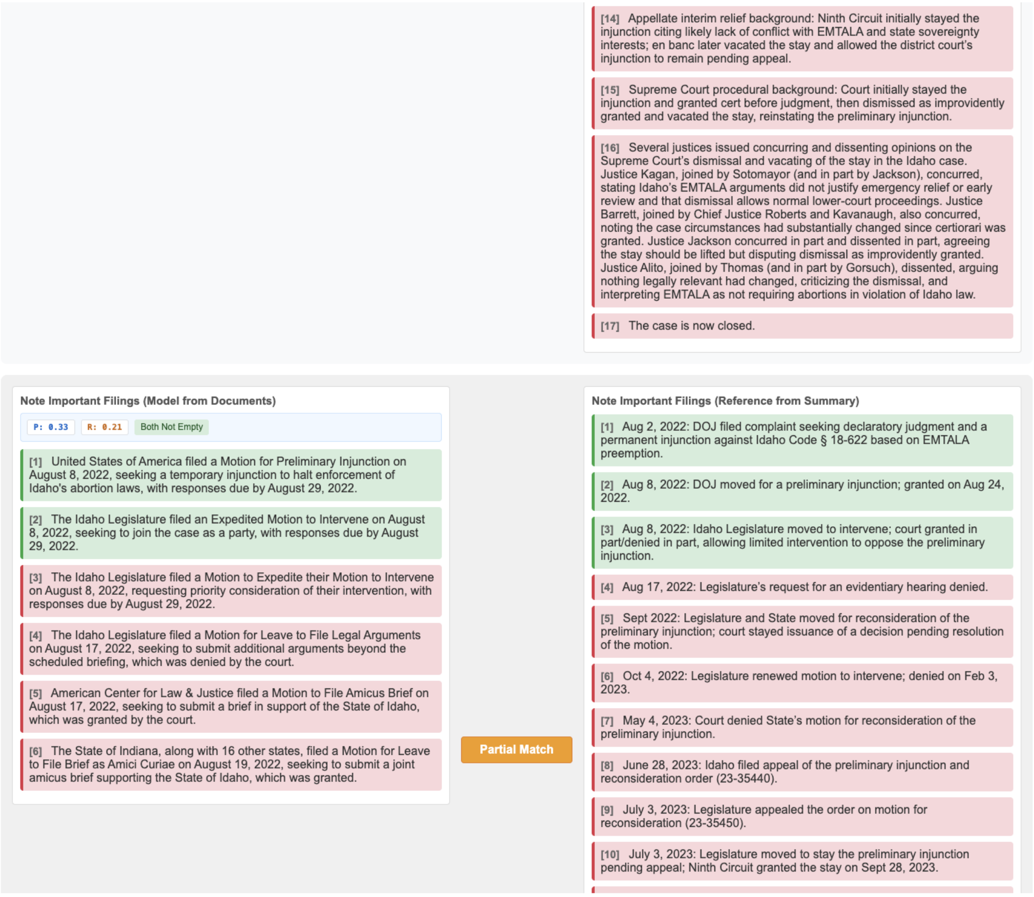
[12] Idaho is barred from prosecuting, disciplining, or imposing any liability on medical providers or hospitals for performing abortions that are necessary to prevent serious jeopardy to a pregnant patient's health, serious impairment of bodily functions, or serious dysfunction of any organ or part.

[13] The court noted that the Idaho Supreme Court's January 2023 decision upholding § 18-622 did not undermine the legal foundations of its own preliminary injunction.

[14] Appellate interim relief background: Ninth Circuit initially stayed the injunction citing likely lack of conflict with EMTALA and state sovereignty interests; en banc later vacated the stay and allowed the district court's injunction to remain pending appeal.

[15] Supreme Court procedural background: Court initially stayed the

Figure 18: Screenshot of a visualization for one case, comparing checklists extracted directly from case documents by GAVEL-AGENT with Qwen3 30B-A3B (26 individual agents configuration) against the human-annotated checklist extracted from the case summary (figure 6 of 10).



[14] Appellate interim relief background: Ninth Circuit initially stayed the injunction citing likely lack of conflict with EMTALA and state sovereignty interests; en banc later vacated the stay and allowed the district court's injunction to remain pending appeal.

[15] Supreme Court procedural background: Court initially stayed the injunction and granted cert before judgment, then dismissed as improvidently granted and vacated the stay, reinstating the preliminary injunction.

[16] Several justices issued concurring and dissenting opinions on the Supreme Court's dismissal and vacating of the stay in the Idaho case. Justice Kagan, joined by Sotomayor (and in part by Jackson), concurred, stating Idaho's EMTALA arguments did not justify emergency relief or early review and that dismissal allows normal lower-court proceedings. Justice Barrett, joined by Chief Justice Roberts and Kavanaugh, also concurred, noting the case circumstances had substantially changed since certiorari was granted. Justice Jackson concurred in part and dissented in part, agreeing the stay should be lifted but disputing dismissal as improvidently granted. Justice Alito, joined by Thomas (and in part by Gorsuch), dissented, arguing nothing legally relevant had changed, criticizing the dismissal, and interpreting EMTALA as not requiring abortions in violation of Idaho law.

[17] The case is now closed.

Note Important Filings (Model from Documents)

[1] United States of America filed a Motion for Preliminary Injunction on August 8, 2022, seeking a temporary injunction to halt enforcement of Idaho's abortion laws, with responses due by August 29, 2022.

[2] The Idaho Legislature filed an Expedited Motion to Intervene on August 8, 2022, seeking to join the case as a party, with responses due by August 29, 2022.

[3] The Idaho Legislature filed a Motion to Expedite their Motion to Intervene on August 8, 2022, requesting priority consideration of their intervention, with responses due by August 29, 2022.

[4] The Idaho Legislature filed a Motion for Leave to File Legal Arguments on August 17, 2022, seeking to submit additional arguments beyond the scheduled briefing, which was denied by the court.

[5] American Center for Law & Justice filed a Motion to File Amicus Brief on August 17, 2022, seeking to submit a brief in support of the State of Idaho, which was granted by the court.

[6] The State of Indiana, along with 16 other states, filed a Motion for Leave to File Brief as Amici Curiae on August 19, 2022, seeking to submit a joint amicus brief supporting the State of Idaho, which was granted.

Note Important Filings (Reference from Summary)

[1] Aug 2, 2022: DOJ filed complaint seeking declaratory judgment and a permanent injunction against Idaho Code § 18-622 based on EMTALA preemption.

[2] Aug 8, 2022: DOJ moved for a preliminary injunction; granted on Aug 24, 2022.

[3] Aug 8, 2022: Idaho Legislature moved to intervene; court granted in part/denied in part, allowing limited intervention to oppose the preliminary injunction.

[4] Aug 17, 2022: Legislature's request for an evidentiary hearing denied.

[5] Sept 2022: Legislature and State moved for reconsideration of the preliminary injunction; court stayed issuance of a decision pending resolution of the motion.

[6] Oct 4, 2022: Legislature renewed motion to intervene; denied on Feb 3, 2023.

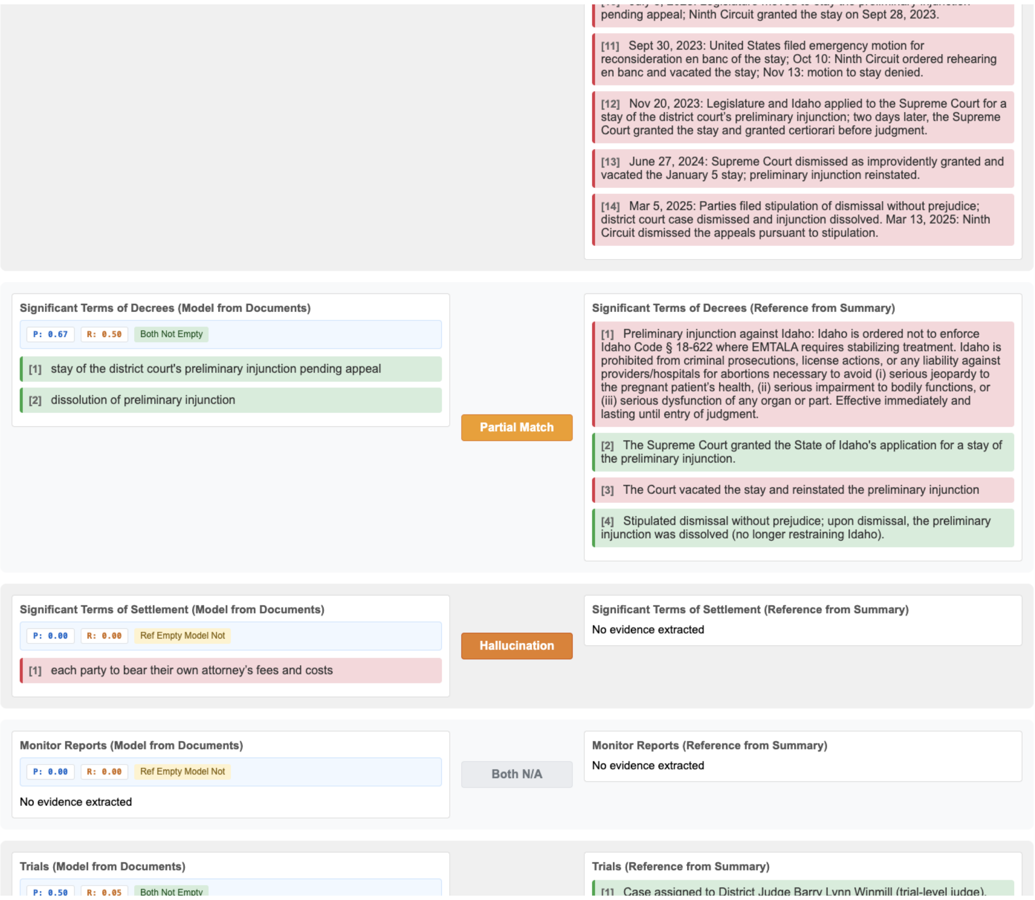
[7] May 4, 2023: Court denied State's motion for reconsideration of the preliminary injunction.

[8] June 28, 2023: Idaho filed appeal of the preliminary injunction and reconsideration order (23-35440).

[9] July 3, 2023: Legislature appealed the order on motion for reconsideration (23-35450).

[10] July 3, 2023: Legislature moved to stay the preliminary injunction pending appeal; Ninth Circuit granted the stay on Sept 28, 2023.

Figure 19: Screenshot of a visualization for one case, comparing checklists extracted directly from case documents by GAVEL-AGENT with Qwen3 30B-A3B (26 individual agents configuration) against the human-annotated checklist extracted from the case summary (figure 7 of 10).



pending appeal; Ninth Circuit granted the stay on Sept 28, 2023.

[11] Sept 30, 2023: United States filed emergency motion for reconsideration en banc of the stay; Oct 10: Ninth Circuit ordered rehearing en banc and vacated the stay; Nov 13: motion to stay denied.

[12] Nov 20, 2023: Legislature and Idaho applied to the Supreme Court for a stay of the district court's preliminary injunction; two days later, the Supreme Court granted the stay and granted certiorari before judgment.

[13] June 27, 2024: Supreme Court dismissed as improvidently granted and vacated the January 5 stay; preliminary injunction reinstated.

[14] Mar 5, 2025: Parties filed stipulation of dismissal without prejudice; district court case dismissed and injunction dissolved. Mar 13, 2025: Ninth Circuit dismissed the appeals pursuant to stipulation.

Significant Terms of Decrees (Model from Documents)

P: 0.67 R: 0.50 Both Not Empty

[1] stay of the district court's preliminary injunction pending appeal

[2] dissolution of preliminary injunction

Partial Match

Significant Terms of Decrees (Reference from Summary)

[1] Preliminary injunction against Idaho: Idaho is ordered not to enforce Idaho Code § 18-622 where EMTALA requires stabilizing treatment. Idaho is prohibited from criminal prosecutions, license actions, or any liability against providers/hospitals for abortions necessary to avoid (i) serious jeopardy to the pregnant patient's health, (ii) serious impairment to bodily functions, or (iii) serious dysfunction of any organ or part. Effective immediately and lasting until entry of judgment.

[2] The Supreme Court granted the State of Idaho's application for a stay of the preliminary injunction.

[3] The Court vacated the stay and reinstated the preliminary injunction

[4] Stipulated dismissal without prejudice; upon dismissal, the preliminary injunction was dissolved (no longer restraining Idaho).

Significant Terms of Settlement (Model from Documents)

P: 0.00 R: 0.00 Ref Empty Model Not

[1] each party to bear their own attorney's fees and costs

Hallucination

Significant Terms of Settlement (Reference from Summary)

No evidence extracted

Monitor Reports (Model from Documents)

P: 0.00 R: 0.00 Ref Empty Model Not

No evidence extracted

Both N/A

Monitor Reports (Reference from Summary)

No evidence extracted

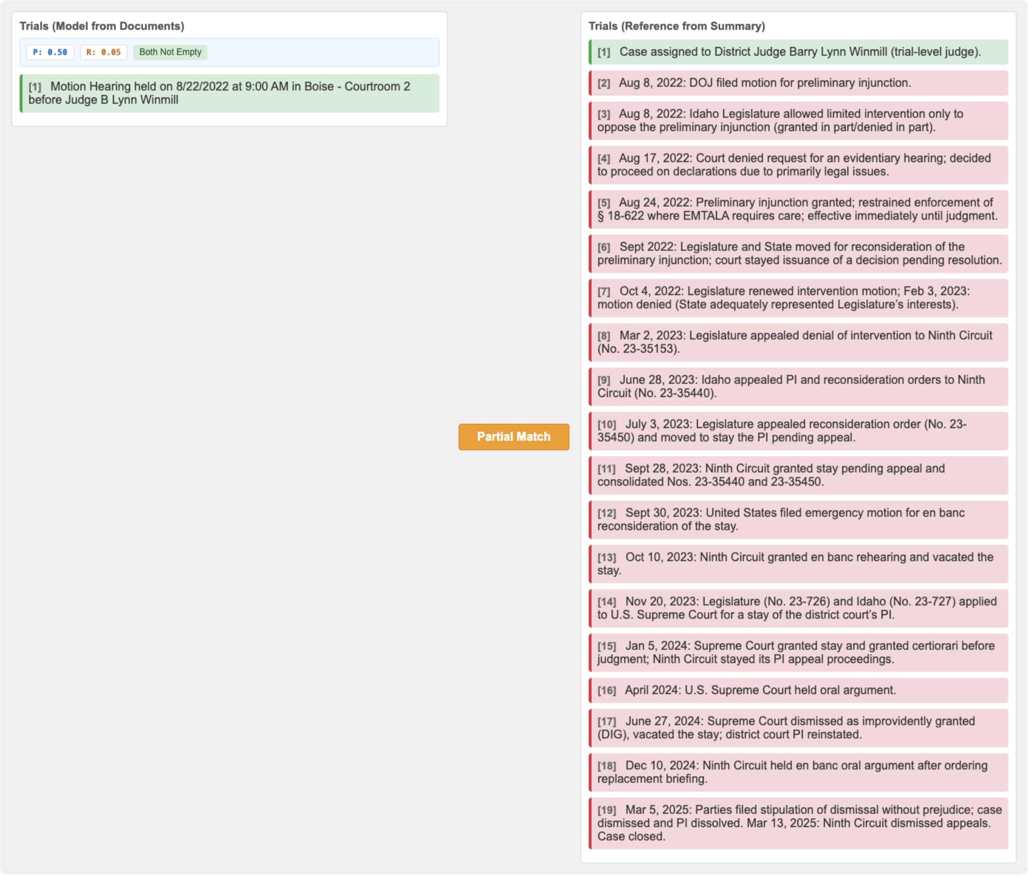
Trials (Model from Documents)

P: 0.50 R: 0.05 Both Not Empty

Trials (Reference from Summary)

[1] Case assigned to District Judge Barry Lynn Winmill (trial-level judge).

Figure 20: Screenshot of a visualization for one case, comparing checklists extracted directly from case documents by GAVEL-AGENT with Qwen3 30B-A3B (26 individual agents configuration) against the human-annotated checklist extracted from the case summary (figure 8 of 10).



Trials (Model from Documents)

P: 0.50 R: 0.05 Both Not Empty

[1] Motion Hearing held on 8/22/2022 at 9:00 AM in Boise - Courtroom 2 before Judge B Lynn Winmill

Trials (Reference from Summary)

[1] Case assigned to District Judge Barry Lynn Winmill (trial-level judge).

[2] Aug 8, 2022: DOJ filed motion for preliminary injunction.

[3] Aug 8, 2022: Idaho Legislature allowed limited intervention only to oppose the preliminary injunction (granted in part/denied in part).

[4] Aug 17, 2022: Court denied request for an evidentiary hearing; decided to proceed on declarations due to primarily legal issues.

[5] Aug 24, 2022: Preliminary injunction granted; restrained enforcement of § 18-622 where EMTALA requires care; effective immediately until judgment.

[6] Sept 2022: Legislature and State moved for reconsideration of the preliminary injunction; court stayed issuance of a decision pending resolution.

[7] Oct 4, 2022: Legislature renewed intervention motion; Feb 3, 2023: motion denied (State adequately represented Legislature's interests).

[8] Mar 2, 2023: Legislature appealed denial of intervention to Ninth Circuit (No. 23-35153).

[9] June 28, 2023: Idaho appealed PI and reconsideration orders to Ninth Circuit (No. 23-35440).

[10] July 3, 2023: Legislature appealed reconsideration order (No. 23-35450) and moved to stay the PI pending appeal.

[11] Sept 28, 2023: Ninth Circuit granted stay pending appeal and consolidated Nos. 23-35440 and 23-35450.

[12] Sept 30, 2023: United States filed emergency motion for en banc reconsideration of the stay.

[13] Oct 10, 2023: Ninth Circuit granted en banc rehearing and vacated the stay.

[14] Nov 20, 2023: Legislature (No. 23-726) and Idaho (No. 23-727) applied to U.S. Supreme Court for a stay of the district court's PI.

[15] Jan 5, 2024: Supreme Court granted stay and granted certiorari before judgment; Ninth Circuit stayed its PI appeal proceedings.

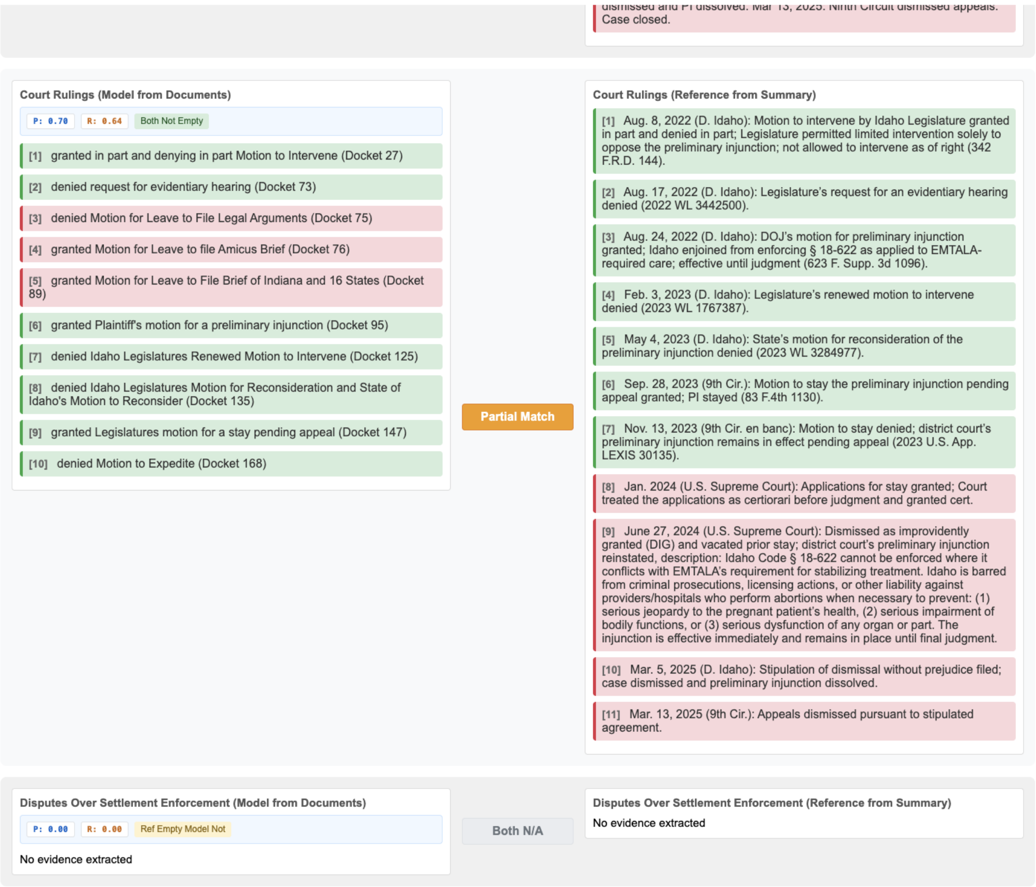
[16] April 2024: U.S. Supreme Court held oral argument.

[17] June 27, 2024: Supreme Court dismissed as improvidently granted (DIG), vacated the stay; district court PI reinstated.

[18] Dec 10, 2024: Ninth Circuit held en banc oral argument after ordering replacement briefing.

[19] Mar 5, 2025: Parties filed stipulation of dismissal without prejudice; case dismissed and PI dissolved. Mar 13, 2025: Ninth Circuit dismissed appeals. Case closed.

Figure 21: Screenshot of a visualization for one case, comparing checklists extracted directly from case documents by GAVEL-AGENT with Qwen3 30B-A3B (26 individual agents configuration) against the human-annotated checklist extracted from the case summary (figure 9 of 10).



dismissed and PI dissolved. Mar 13, 2025: Ninth Circuit dismissed appeals. Case closed.

Court Rulings (Model from Documents)

P: 0.70 R: 0.64 Both Not Empty

[1] granted in part and denying in part Motion to Intervene (Docket 27)

[2] denied request for evidentiary hearing (Docket 73)

[3] denied Motion for Leave to File Legal Arguments (Docket 75)

[4] granted Motion for Leave to file Amicus Brief (Docket 76)

[5] granted Motion for Leave to File Brief of Indiana and 16 States (Docket 89)

[6] granted Plaintiff's motion for a preliminary injunction (Docket 95)

[7] denied Idaho Legislatures Renewed Motion to Intervene (Docket 125)

[8] denied Idaho Legislatures Motion for Reconsideration and State of Idaho's Motion to Reconsider (Docket 135)

[9] granted Legislatures motion for a stay pending appeal (Docket 147)

[10] denied Motion to Expedite (Docket 168)

Court Rulings (Reference from Summary)

[1] Aug. 8, 2022 (D. Idaho): Motion to intervene by Idaho Legislature granted in part and denied in part; Legislature permitted limited intervention solely to oppose the preliminary injunction; not allowed to intervene as of right (342 F.R.D. 144).

[2] Aug. 17, 2022 (D. Idaho): Legislature's request for an evidentiary hearing denied (2022 WL 3442500).

[3] Aug. 24, 2022 (D. Idaho): DOJ's motion for preliminary injunction granted; Idaho enjoined from enforcing § 18-622 as applied to EMTALA-required care; effective until judgment (623 F. Supp. 3d 1096).

[4] Feb. 3, 2023 (D. Idaho): Legislature's renewed motion to intervene denied (2023 WL 1767387).

[5] May 4, 2023 (D. Idaho): State's motion for reconsideration of the preliminary injunction denied (2023 WL 3284977).

[6] Sep. 28, 2023 (9th Cir.): Motion to stay the preliminary injunction pending appeal granted; PI stayed (83 F.4th 1130).

[7] Nov. 13, 2023 (9th Cir. en banc): Motion to stay denied; district court's preliminary injunction remains in effect pending appeal (2023 U.S. App. LEXIS 30135).

[8] Jan. 2024 (U.S. Supreme Court): Applications for stay granted; Court treated the applications as certiorari before judgment and granted cert.

[9] June 27, 2024 (U.S. Supreme Court): Dismissed as improvidently granted (DIG) and vacated prior stay; district court's preliminary injunction reinstated, description: Idaho Code § 18-622 cannot be enforced where it conflicts with EMTALA's requirement for stabilizing treatment. Idaho is barred from criminal prosecutions, licensing actions, or other liability against providers/hospitals who perform abortions when necessary to prevent: (1) serious jeopardy to the pregnant patient's health, (2) serious impairment of bodily functions, or (3) serious dysfunction of any organ or part. The injunction is effective immediately and remains in place until final judgment.

[10] Mar. 5, 2025 (D. Idaho): Stipulation of dismissal without prejudice filed; case dismissed and preliminary injunction dissolved.

[11] Mar. 13, 2025 (9th Cir.): Appeals dismissed pursuant to stipulated agreement.

Disputes Over Settlement Enforcement (Model from Documents)

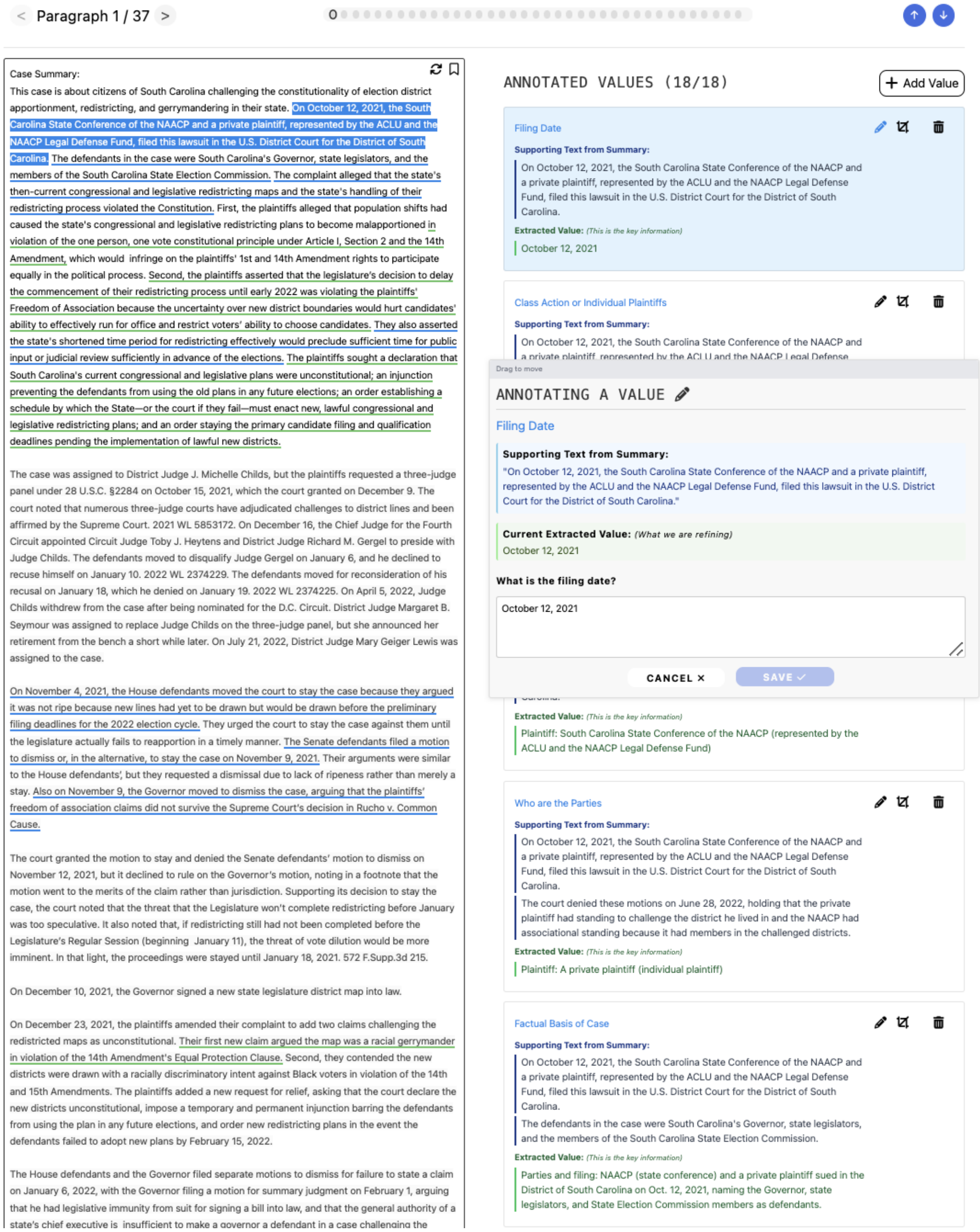
P: 0.00 R: 0.00 Ref Empty Model Not

No evidence extracted

Disputes Over Settlement Enforcement (Reference from Summary)

No evidence extracted

Figure 22: Screenshot of a visualization for one case, comparing checklists extracted directly from case documents by GAVEL-AGENT with Qwen3 30B-A3B (26 individual agents configuration) against the human-annotated checklist extracted from the case summary (figure 10 of 10).



Case Summary:

This case is about citizens of South Carolina challenging the constitutionality of election district apportionment, redistricting, and gerrymandering in their state. On October 12, 2021, the South Carolina State Conference of the NAACP and a private plaintiff, represented by the ACLU and the NAACP Legal Defense Fund, filed this lawsuit in the U.S. District Court for the District of South Carolina. The defendants in the case were South Carolina's Governor, state legislators, and the members of the South Carolina State Election Commission. The complaint alleged that the state's then-current congressional and legislative redistricting maps and the state's handling of their redistricting process violated the Constitution. First, the plaintiffs alleged that population shifts had caused the state's congressional and legislative redistricting plans to become malapportioned in violation of the one person, one vote constitutional principle under Article I, Section 2 and the 14th Amendment, which would infringe on the plaintiffs' 1st and 14th Amendment rights to participate equally in the political process. Second, the plaintiffs asserted that the legislature's decision to delay the commencement of their redistricting process until early 2022 was violating the plaintiffs' Freedom of Association because the uncertainty over new district boundaries would hurt candidates' ability to effectively run for office and restrict voters' ability to choose candidates. They also asserted the state's shortened time period for redistricting effectively would preclude sufficient time for public input or judicial review sufficiently in advance of the elections. The plaintiffs sought a declaration that South Carolina's current congressional and legislative plans were unconstitutional; an injunction preventing the defendants from using the old plans in any future elections; an order establishing a schedule by which the State—or the court if they fail—must enact new, lawful congressional and legislative redistricting plans; and an order staying the primary candidate filing and qualification deadlines pending the implementation of lawful new districts.

The case was assigned to District Judge J. Michelle Childs, but the plaintiffs requested a three-judge panel under 28 U.S.C. §2284 on October 15, 2021, which the court granted on December 9. The court noted that numerous three-judge courts have adjudicated challenges to district lines and been affirmed by the Supreme Court. 2021 WL 5853172. On December 16, the Chief Judge for the Fourth Circuit appointed Circuit Judge Toby J. Heytens and District Judge Richard M. Gergel to preside with Judge Childs. The defendants moved to disqualify Judge Gergel on January 6, and he declined to recuse himself on January 10. 2022 WL 2374229. The defendants moved for reconsideration of his recusal on January 18, which he denied on January 19. 2022 WL 2374225. On April 5, 2022, Judge Childs withdrew from the case after being nominated for the D.C. Circuit. District Judge Margaret B. Seymour was assigned to replace Judge Childs on the three-judge panel, but she announced her retirement from the bench a short while later. On July 21, 2022, District Judge Mary Geiger Lewis was assigned to the case.

On November 4, 2021, the House defendants moved the court to stay the case because they argued it was not ripe because new lines had yet to be drawn but would be drawn before the preliminary filing deadlines for the 2022 election cycle. They urged the court to stay the case against them until the legislature actually fails to reapportion in a timely manner. The Senate defendants filed a motion to dismiss or, in the alternative, to stay the case on November 9, 2021. Their arguments were similar to the House defendants', but they requested a dismissal due to lack of ripeness rather than merely a stay. Also on November 9, the Governor moved to dismiss the case, arguing that the plaintiffs' freedom of association claims did not survive the Supreme Court's decision in Rucho v. Common Cause.

The court granted the motion to stay and denied the Senate defendants' motion to dismiss on November 12, 2021, but it declined to rule on the Governor's motion, noting in a footnote that the motion went to the merits of the claim rather than jurisdiction. Supporting its decision to stay the case, the court noted that the threat that the Legislature won't complete redistricting before January was too speculative. It also noted that, if redistricting still had not been completed before the Legislature's Regular Session (beginning January 11), the threat of vote dilution would be more imminent. In that light, the proceedings were stayed until January 18, 2021. 572 F.Supp.3d 215.

On December 10, 2021, the Governor signed a new state legislature district map into law.

On December 23, 2021, the plaintiffs amended their complaint to add two claims challenging the redistricted maps as unconstitutional. Their first new claim argued the map was a racial gerrymander in violation of the 14th Amendment's Equal Protection Clause. Second, they contended the new districts were drawn with a racially discriminatory intent against Black voters in violation of the 14th and 15th Amendments. The plaintiffs added a new request for relief, asking that the court declare the new districts unconstitutional, impose a temporary and permanent injunction barring the defendants from using the plan in any future elections, and order new redistricting plans in the event the defendants failed to adopt new plans by February 15, 2022.

The House defendants and the Governor filed separate motions to dismiss for failure to state a claim on January 6, 2022, with the Governor filing a motion for summary judgment on February 1, arguing that he had legislative immunity from suit for signing a bill into law, and that the general authority of a state's chief executive is insufficient to make a governor a defendant in a case challenging the

ANNOTATED VALUES (18/18)

Filing Date

Supporting Text from Summary:

On October 12, 2021, the South Carolina State Conference of the NAACP and a private plaintiff, represented by the ACLU and the NAACP Legal Defense Fund, filed this lawsuit in the U.S. District Court for the District of South Carolina.

Extracted Value: *(This is the key information)*

October 12, 2021

Class Action or Individual Plaintiffs

Supporting Text from Summary:

On October 12, 2021, the South Carolina State Conference of the NAACP and a private plaintiff, represented by the ACLU and the NAACP Legal Defense

ANNOTATING A VALUE

Filing Date

Supporting Text from Summary:

"On October 12, 2021, the South Carolina State Conference of the NAACP and a private plaintiff, represented by the ACLU and the NAACP Legal Defense Fund, filed this lawsuit in the U.S. District Court for the District of South Carolina."

Current Extracted Value: *(What we are refining)*

October 12, 2021

What is the filing date?

October 12, 2021

Extracted Value: *(This is the key information)*

Plaintiff: South Carolina State Conference of the NAACP (represented by the ACLU and the NAACP Legal Defense Fund)

Who are the Parties

Supporting Text from Summary:

On October 12, 2021, the South Carolina State Conference of the NAACP and a private plaintiff, represented by the ACLU and the NAACP Legal Defense Fund, filed this lawsuit in the U.S. District Court for the District of South Carolina.

The court denied these motions on June 28, 2022, holding that the private plaintiff had standing to challenge the district he lived in and the NAACP had associational standing because it had members in the challenged districts.

Extracted Value: *(This is the key information)*

Plaintiff: A private plaintiff (individual plaintiff)

Factual Basis of Case

Supporting Text from Summary:

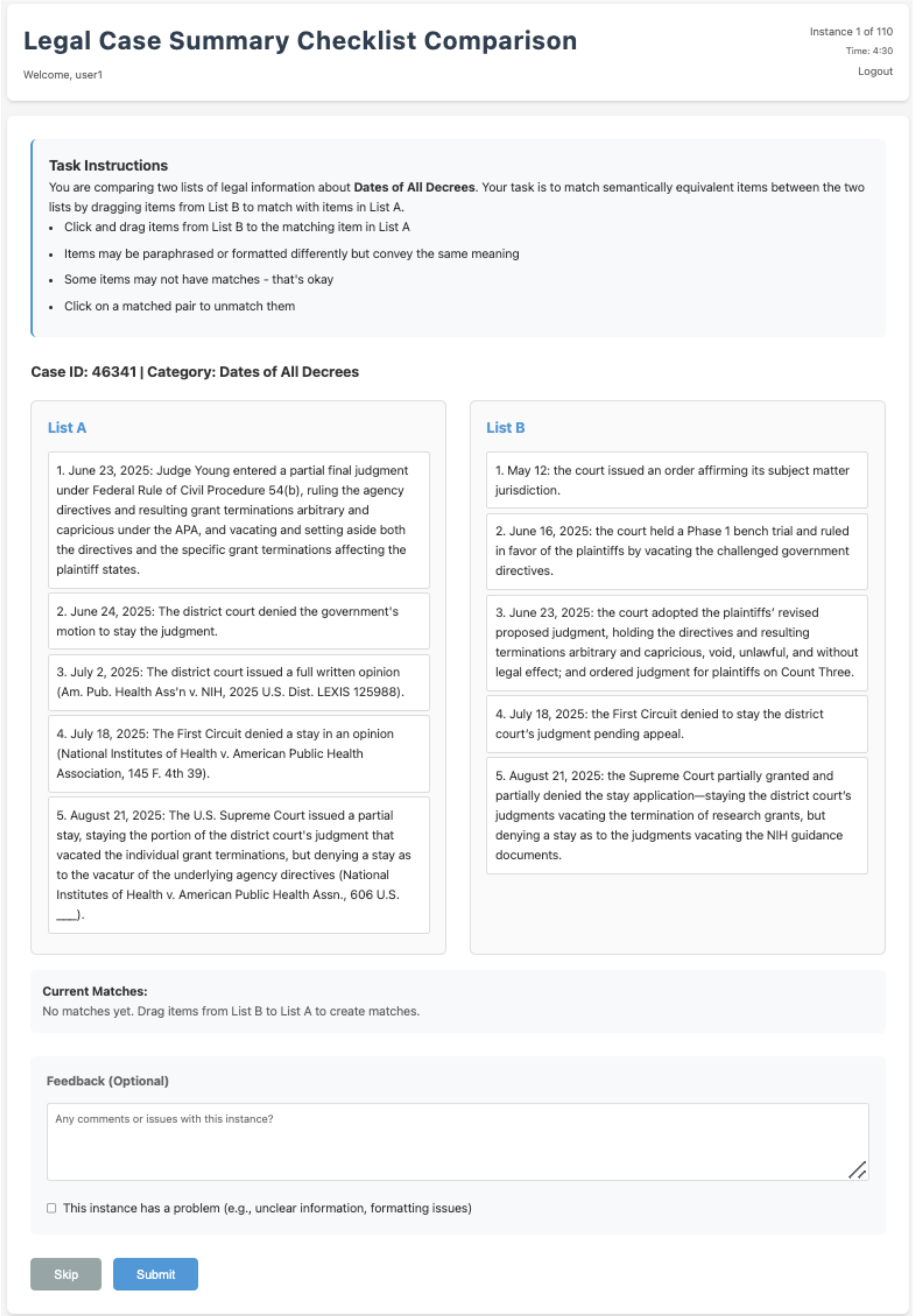
On October 12, 2021, the South Carolina State Conference of the NAACP and a private plaintiff, represented by the ACLU and the NAACP Legal Defense Fund, filed this lawsuit in the U.S. District Court for the District of South Carolina.

The defendants in the case were South Carolina's Governor, state legislators, and the members of the South Carolina State Election Commission.

Extracted Value: *(This is the key information)*

Parties and filing: NAACP (state conference) and a private plaintiff sued in the District of South Carolina on Oct. 12, 2021, naming the Governor, state legislators, and State Election Commission members as defendants.

Figure 23: Screenshot of the annotation interface for checklist extraction from summaries. Annotators can add, remove, or modify checklist item values, with the process carried out paragraph by paragraph to ensure each sentence is carefully reviewed.



Legal Case Summary Checklist Comparison

Welcome, user1

Instance 1 of 110
Time: 4:30

Task Instructions

You are comparing two lists of legal information about **Dates of All Decrees**. Your task is to match semantically equivalent items between the two lists by dragging items from List B to match with items in List A.

- Click and drag items from List B to the matching item in List A
- Items may be paraphrased or formatted differently but convey the same meaning
- Some items may not have matches - that's okay
- Click on a matched pair to unmatch them

Case ID: 46341 | Category: Dates of All Decrees

List A

1. June 23, 2025: Judge Young entered a partial final judgment under Federal Rule of Civil Procedure 54(b), ruling the agency directives and resulting grant terminations arbitrary and capricious under the APA, and vacating and setting aside both the directives and the specific grant terminations affecting the plaintiff states.

2. June 24, 2025: The district court denied the government's motion to stay the judgment.

3. July 2, 2025: The district court issued a full written opinion (Am. Pub. Health Ass'n v. NIH, 2025 U.S. Dist. LEXIS 125988).

4. July 18, 2025: The First Circuit denied a stay in an opinion (National Institutes of Health v. American Public Health Association, 145 F. 4th 39).

5. August 21, 2025: The U.S. Supreme Court issued a partial stay, staying the portion of the district court's judgment that vacated the individual grant terminations, but denying a stay as to the vacatur of the underlying agency directives (National Institutes of Health v. American Public Health Assn., 606 U.S. ___).

List B

1. May 12: the court issued an order affirming its subject matter jurisdiction.

2. June 16, 2025: the court held a Phase 1 bench trial and ruled in favor of the plaintiffs by vacating the challenged government directives.

3. June 23, 2025: the court adopted the plaintiffs' revised proposed judgment, holding the directives and resulting terminations arbitrary and capricious, void, unlawful, and without legal effect; and ordered judgment for plaintiffs on Count Three.

4. July 18, 2025: the First Circuit denied to stay the district court's judgment pending appeal.

5. August 21, 2025: the Supreme Court partially granted and partially denied the stay application—staying the district court's judgments vacating the termination of research grants, but denying a stay as to the judgments vacating the NIH guidance documents.

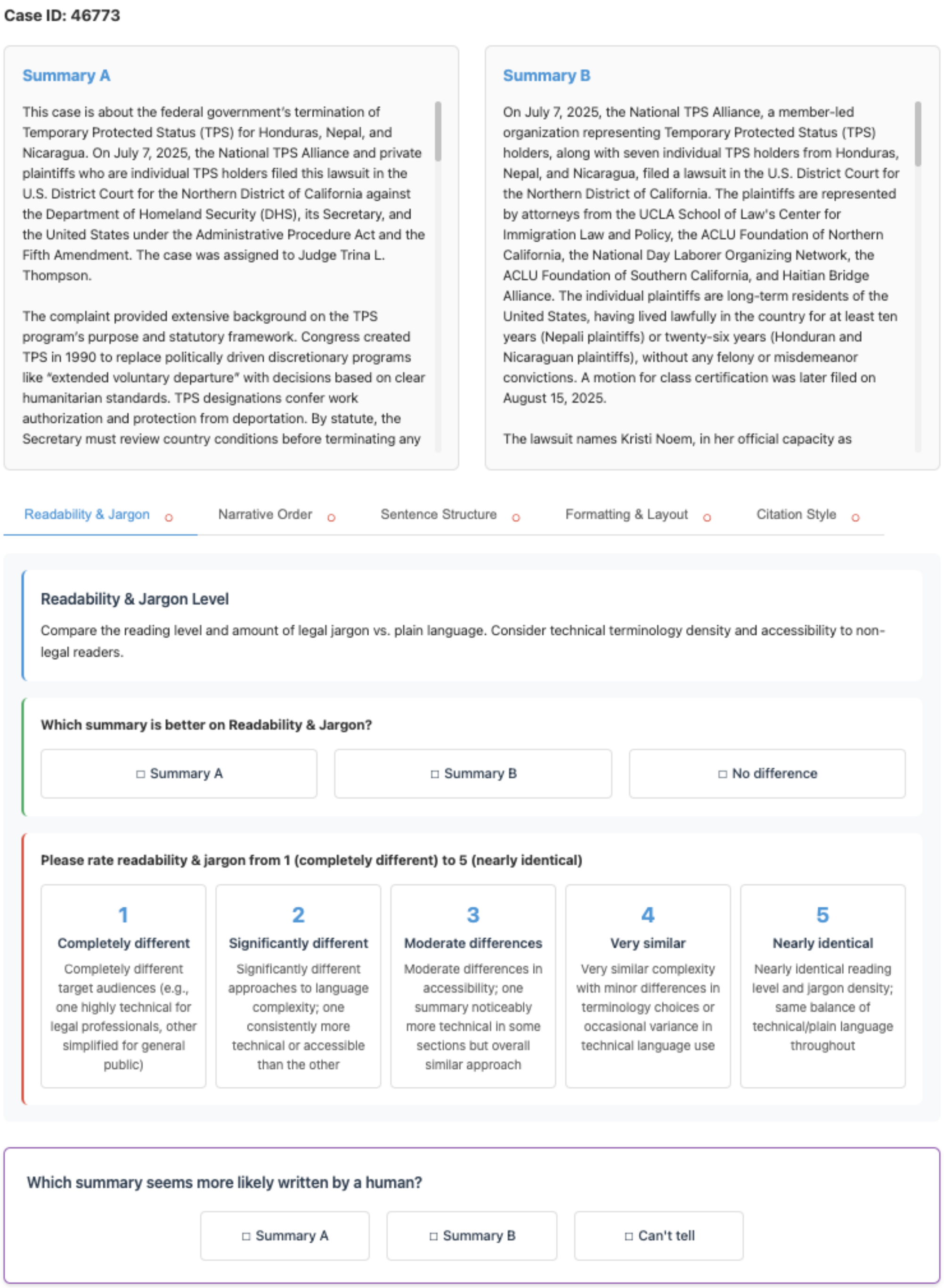
Current Matches:
No matches yet. Drag items from List B to List A to create matches.

Feedback (Optional)

Any comments or issues with this instance?

□ This instance has a problem (e.g., unclear information, formatting issues)

Figure 24: Screenshot of the annotation interface for checklist comparison. Annotators match items between two lists in a list-wise comparison. For string-wise comparison, where both values are strings, the middle component becomes a radio selection with four options: equal, A contains B, B contains A, or different.



Case ID: 46773

Summary A

This case is about the federal government's termination of Temporary Protected Status (TPS) for Honduras, Nepal, and Nicaragua. On July 7, 2025, the National TPS Alliance and private plaintiffs who are individual TPS holders filed this lawsuit in the U.S. District Court for the Northern District of California against the Department of Homeland Security (DHS), its Secretary, and the United States under the Administrative Procedure Act and the Fifth Amendment. The case was assigned to Judge Trina L. Thompson.

The complaint provided extensive background on the TPS program's purpose and statutory framework. Congress created TPS in 1990 to replace politically driven discretionary programs like "extended voluntary departure" with decisions based on clear humanitarian standards. TPS designations confer work authorization and protection from deportation. By statute, the Secretary must review country conditions before terminating any

Summary B

On July 7, 2025, the National TPS Alliance, a member-led organization representing Temporary Protected Status (TPS) holders, along with seven individual TPS holders from Honduras, Nepal, and Nicaragua, filed a lawsuit in the U.S. District Court for the Northern District of California. The plaintiffs are represented by attorneys from the UCLA School of Law's Center for Immigration Law and Policy, the ACLU Foundation of Northern California, the National Day Laborer Organizing Network, the ACLU Foundation of Southern California, and Haitian Bridge Alliance. The individual plaintiffs are long-term residents of the United States, having lived lawfully in the country for at least ten years (Nepali plaintiffs) or twenty-six years (Honduran and Nicaraguan plaintiffs), without any felony or misdemeanor convictions. A motion for class certification was later filed on August 15, 2025.

The lawsuit names Kristi Noem, in her official capacity as

Readability & Jargon Level

Compare the reading level and amount of legal jargon vs. plain language. Consider technical terminology density and accessibility to non-legal readers.

Which summary is better on Readability & Jargon?

□ Summary A □ Summary B □ No difference

Please rate readability & jargon from 1 (completely different) to 5 (nearly identical)

1
Completely different
Completely different target audiences (e.g., one highly technical for legal professionals, other simplified for general public)

2
Significantly different
Significantly different approaches to language complexity; one consistently more technical or accessible than the other

3
Moderate differences
Moderate differences in accessibility; one summary noticeably more technical in some sections but overall similar approach

4
Very similar
Very similar complexity with minor differences in terminology choices or occasional variance in technical language use

5
Nearly identical
Nearly identical reading level and jargon density; same balance of technical/plain language throughout

Which summary seems more likely written by a human?

□ Summary A □ Summary B □ Can't tell

Figure 25: Screenshot of the annotation interface for rating writing style similarity. Annotators compare two summaries, providing ratings on five aspects and answering auxiliary questions such as which summary they prefer.

Prompt for Extracting Checklist from Summary

```
You are assisting a lawyer in extracting key information from a
↪  legal case summary. Given a case summary, identify
↪  {checklist_item_definition}
# Note: Do not make assumptions or add information that is not
↪  presented in the summary.

# Case Summary
{case_summary}

# Output Format
Your output should be in the following JSON format-no extra keys,
↪  no prose outside of the JSON:

```
{{
  "reasoning": "<brief analysis of the case summary and how you
  ↪ identified the relevant information or determined that none
  ↪ was present>",
  "extracted": [
    {{
      "evidence": [
        "<verbatim snippet 1>",
        "<verbatim snippet 2 (if multiple snippets are relevant)>"
        // ...
      ],
      "value": "<extracted information from the evidence>"
    }}
    // ...
  ]
}}
```
## Definitions of each part
- `reasoning`: A brief analysis of the case summary and how you
↪  identified the relevant information or determined that none was
↪  present.
- `extracted`: A list of one or more objects, each representing a
↪  distinct piece of information relevant to the checklist item
↪  (e.g., multiple court rulings, decree dates, or cited
↪  opinions). Always use a list, even if there is only one item.
- `evidence`: One or more exact text snippets copied from the case
↪  summary that support the extracted information. Always return
↪  as a list of strings.
- `value`: The extracted information.

## Rules for the JSON schema
1. **extracted** and **evidence** is always a list, even if they
↪  hold a single object.
2. Copy the **evidence** exactly as it appears in the case
↪  summary-no rewriting.
3. If the case summary contains no relevant information, output the
↪  **extracted** as an empty list:

```
{{
  "reasoning": "<brief analysis>",
  "extracted": []
}}
```
```

Figure 26

Prompt for Comparing Single-Value Checklist Item

```
You are given two pieces of legal information (A and B) about
 ↪  **{checklist_category}**, extracted from two summaries of
 ↪  the same case. Your task is to compare these pieces of
 ↪  information based on their **semantic meaning** - that
 ↪  is, what they actually convey, regardless of how they are
 ↪  worded or formatted.

# Information to Compare
## Information A:
{information_A}

## Information B:
{information_B}

# Relationship Options
Determine which of these four relationships best describes
 ↪  how A and B relate to each other:
1. **"A contains B"** - A includes all the information in B,
 ↪  plus additional information
2. **"B contains A"** - B includes all the information in A,
 ↪  plus additional information
3. **"A equals B"** - A and B convey the same information
 ↪  (semantically equivalent)
4. **"A and B are different"** - A and B contain different or
 ↪  conflicting information

# Output Format
Structure your response as follows:
**Reasoning:** Provide your detailed analysis of how the two
 ↪  pieces of information relate to each other

**Final Answer:** State one of the four options: "A contains
 ↪  B", "B contains A", "A equals B", or "A and B are
 ↪  different"
```

Figure 27

Prompt for Comparing Multi-Value Checklist Item

~~~
You are given two lists of legal information (A and B) about **{checklist_category}**, extracted from two summaries of the same legal case. Your task is to compare these lists based on their **semantic meaning**-that is, what each item conveys, regardless of wording, format, or phrasing.

You should identify:
1. Items that appear in **both A and B** (i.e., semantically equivalent),
2. Items that appear **only in A**,
3. Items that appear **only in B**.

# Information to Compare
## List A:
{information_A}

## List B:
{information_B}

# Output Format
Structure your response as follows:
**Reasoning:**
Provide your detailed analysis of how the two lists relate to each other. Explain any mappings between items, and how you determined whether they were equivalent or different.

**Final Answer:**
Output a valid JSON object with the following structure:

```json
{{
  "common": [
    {{"A_index": X, "B_index": Y}},
    ...
  ],
  "only_in_A": [X, ...],
  "only_in_B": [Y, ...]
}}
```

Where:
- `A_index` is the index of the item in List A,
- `B_index` is the index of the semantically equivalent item in List B,
- `only_in_A` lists the indices of items in A that do **not** appear in B,
- `only_in_B` lists the indices of items in B that do **not** appear in A.

# Notes
- Both List A and B are numbered using 1-based indexing.
- Match items even if they are paraphrased or formatted differently.
- Treat legal synonyms and abbreviations as equivalent when appropriate.
- Return only valid JSON in the **Final Answer** section.
~~~

Figure 28

Prompt for Extract Residual Facts from Uncovered Text by the Checklist Items

```
You are assisting a lawyer in identifying key information from a
↪  legal case summary. You will be given a set of text spans
↪  extracted from the summary that may contain meaningful legal or
↪  factual content.

Your task is to extract distinct atomic facts from the given spans.
↪  Each atomic fact should be a single discrete, self-contained,
↪  and verifiable piece of information that can stand on its own.
↪  Ignore any spans that contain filler phrases, incomplete
↪  clauses, or do not convey meaningful information. If multiple
↪  spans express the same fact, extract it only once.

# Note: Do not make assumptions or add information that is not
↪  present in the spans.

# Text Spans
{text_spans}

# Output Format

Your output should be in the following JSON format-no extra keys,
↪  no prose outside of the JSON:

```
{{
  "reasoning": "<brief analysis of which spans contain meaningful
  ↪  factual information and what those facts are>",
  "extracted": [
    {{
      "fact": "<atomic fact 1>",
      "evidence_spans": [<list of 1-based span indices>]
    }},
    {{
      "fact": "<atomic fact 2>",
      "evidence_spans": [<list of 1-based span indices>]
    }}
    // ...
  ]
}}
```


## Definitions of each part
* `reasoning`: A brief analysis of the spans and how you identified
↪ any meaningful atomic facts.
* `extracted`: A list of objects, each representing one atomic fact.
↪ Every object must have:
  - `fact`: A clear, concise sentence or phrase conveying a
  ↪ distinct, self-contained fact.
  - `evidence_spans`: A list of 1-based indices of the spans that
  ↪ support or directly contain the fact.

## Rules for the JSON schema
{it is the same as the checklist extraction prompt.}
```

Figure 29

Prompt for Rating Writing Style Similarity on Five Aspects

~~~
You are given two summaries of the same legal case (Summary A and
↪  Summary B). Your task is to evaluate how similar they are in
↪  terms of structure and writing style across five specific
↪  dimensions. You should focus on **similarity** rather than
↪  quality-we want to know how alike these summaries are, not
↪  which one is better.

# Summaries to Compare
## Summary A:
{summary_A}

## Summary B:
{summary_B}

# Evaluation Dimensions with Specific Similarity Scales

{all_5_aspects_definitions}

# Output Format

Structure your response as follows:

**Analysis:**
Provide a detailed comparison for each dimension, explaining
↪  specific similarities and differences you observe between
↪  Summary A and Summary B.

**Scores:**
Output a valid JSON object with your similarity ratings:

```json
{{
  "readability_jargon": X,
  "narrative_order": X,
  "sentence_structure": X,
  "formatting_layout": X,
  "citation_style": X
}}
```

Where X is your similarity rating (1-5) for each dimension.

# Important Notes
- Focus on similarity, not quality or factual correctness
- Evaluate style and structure only, ignore content differences
- Consider the summaries as a whole when rating each dimension
- Apply the scale objectively for every dimension, strictly
↪  following each definition
~~~

Figure 30

Prompt for Legal Summarization

```
You are given multiple documents related to a legal case. Your task
 ↪  is to generate a clear, legally precise, and self-contained
 ↪  summary that would let the reader grasp the case without
 ↪  consulting the source files without being excessively long or
 ↪  overly detailed.

Write the summary as a factual narrative. The checklist below shows
 ↪  what to include. Items marked "(if applicable)" should only be
 ↪  included when relevant. If information isn't in the documents,
 ↪  omit it-do not speculate.

# Legal Case Summary Checklist
{all_26_checklist_item_definitions}

# Case Documents
{case_documents}

# Output Format
Please structure your response as follows:
**Reasoning:** Briefly explain what key elements you focused on in
 ↪  the documents to build your summary.

**Case Summary:** A clear, legally precise narrative of the case,
 ↪  written in paragraph form, without being too long.

# Guidelines
* Write as a narrative in paragraph form using clear language. Use
 ↪  a logical order-chronological if helpful, but flexible if
 ↪  another sequence improves clarity.
* Include enough detail for understanding while remaining concise.
* Use accurate legal terminology but avoid jargon-write for a
 ↪  general audience.
* Stay strictly factual; do not add analysis beyond what appears in
 ↪  the record.

Now read the case documents and generate the summary following the
 ↪  checklist, output format, and guidelines above.
```

Figure 31

Prompt for End-to-End Extracting Checklist Item from Case Document (Part 1/2)

```
You are assisting a lawyer in extracting key information from legal
 ↪  case documents. You will be given multiple documents related to
 ↪  a legal case. Your task is to {item_description}

# Note:
- Do not make assumptions or add information that is not presented
 ↪  in the documents.
- When extracting evidence, quote the exact text from the
 ↪  documents.
- Each extracted value must be self-contained and easy to
 ↪  understand; include important context when available.

# Case Documents
{case_documents}

# Output Format
Your output should be in the following JSON format-no extra keys,
 ↪  no prose outside of the JSON:

```
{
  "reasoning": "<brief analysis of the case documents and how you
   ↪ identified the relevant information or determined that none
   ↪ was present>",
  "extracted": [
    {
      "evidence": [
        {
          "text": "<verbatim snippet 1>",
          "source_document": "<document name>",
          "location": "<page number or section>"
        },
        {
          "text": "<verbatim snippet 2 (if multiple snippets are
           ↪ relevant)>",
          "source_document": "<document name>",
          "location": "<page number or section>"
        }
        // ...
      ],
      "value": "<extracted information from the evidence>"
    }
    // ...
  ]
}
```
```

Figure 32

Prompt for End-to-End Extracting Checklist Item from Case Document (Part 2/2)

~~~
## Definitions of each part
- `reasoning`: A brief analysis of the case documents and how you identified the relevant information or determined that none was present.
- `extracted`: A list of one or more objects, each representing a distinct piece of information relevant to the checklist item. Always use a list, even if there is only one item.
- `evidence`: A list of evidence objects, each containing:
  - `text`: Exact text snippet copied from the case documents
  - `source_document`: The title/name of the document where this evidence was found
  - `location`: The page number or section identifier where the evidence appears
- `value`: The extracted information based on the evidence.

## Rules for the JSON schema
1. **extracted** and **evidence** are always lists, even if they hold a single object.
2. Copy the **text** in evidence objects exactly as it appears in the case documents-no rewriting or paraphrasing.
3. Always include **source_document** and **location** for each piece of evidence.
4. If the case documents contain no relevant information, output the **extracted** as an empty list:

```
{
  "reasoning": "<brief analysis>",
  "extracted": []
}
```

5. Extract information from all relevant documents-do not stop after finding information in just one document.
6. Each distinct piece of information should be a separate item in the **extracted** list.
7. If you cannot determine the specific page number or section, you may use descriptive locations like "beginning of document", "middle section", or "near the end".
~~~

Figure 33

Prompt for Chunk-by-Chunk Extracting Checklist Items from Case Documents

~~~
You are assisting a lawyer in extracting key information from legal
↪ case documents. You will be given a document chunk from a legal
↪ case. Your task is to {item_description}

# Note:
{same as the end-to-end prompt}

# Current State
This is the accumulated extraction state from previous chunks:
{current_state}

# Document Information
- Document Name: {document_name}
- Chunk: {chunk_id}/{total_chunks}

# Document Chunk
{document_chunk}

# Output Format
Your output should be in the following JSON format-no extra keys,
↪ no prose outside of the JSON:

```
{{
  "reasoning": "<brief analysis of this document chunk and how you
  ↪ identified any new relevant information or determined that
  ↪ none was present>",
  "extracted": [
    {{
      "evidence": [
        {{
          "text": "<verbatim snippet 1>",
          "source_document": "<document name>",
          "location": "Chunk {chunk_id}/{total_chunks}"
        }},
        {{
          "text": "<verbatim snippet 2 (if multiple snippets are
          ↪ relevant)>",
          "source_document": "<document name>",
          "location": "Chunk {chunk_id}/{total_chunks}"
        }}
        // ...
      ],
      "value": "<extracted information from the evidence>"
    }}
    // ...
  ]
}}
```

## Definitions of each part
{same as the end-to-end prompt}

## Rules for the JSON schema
{{same as the end-to-end prompt}}
~~~

Figure 34

System Prompt used in GAVEL-AGENT (Part 1/3)

```
You are a document extraction specialist. Your task is to extract
 ↪  **all checklist items specified in the snapshot** from the
 ↪  provided documents, citing evidence for every value.

  You operate by analyzing the snapshot and selecting **exactly ONE
   ↪  action per turn**. You must **respond with valid JSON only**
   ↪  - no prose, no extra keys.

  # Snapshot
  Provided every turn:
  - Task description
  - Checklist definitions (what items to extract; any number of
   ↪  items)
  - Document catalog with coverage statistics (and
   ↪  catalog_state/version)
  - Checklist summary (which keys are filled/empty/Not Applicable)
  - Recent action history

  # Goal
  Systematically extract all applicable checklist items with proper
   ↪  evidence.

  # Decision Policy
  Choose exactly one action each turn:
  - If the document catalog is **unknown** -> call `list_documents`.
  - If a specific document likely contains a target value, choose
   ↪  ONE:
    * `read_document` - default choice. Read a targeted window
     ↪  (<=10,000 tokens) in a document.
    * `search_document_regex` - use this when the target is clearly
     ↪  patternable (e.g., "Case No.", "Filed:", citations).
  - When you have confirmed text for one or more keys:
    - Use `append_checklist` for adds new entries for some checklist
     ↪  items.
    - Use `update_checklist` to replace the entire extracted list
     ↪  for some checklist items when you have the
     ↪  authoritative/complete set, when correcting earlier
     ↪  entries, or when setting an item to Not Applicable (see
     ↪  "Not Applicable Encoding").
  - Periodically use `get_checklist` to assess remaining gaps.
  - Stop when all keys are filled or set to Not Applicable.

  # Systematic Extraction Process
  **After each read_document or search_document_regex action:**
  - Carefully analyze the returned text to identify ALL checklist
   ↪  items that can be extracted.
  - Cross-reference the text against your checklist definitions to
   ↪  avoid missing relevant values.
  - Your next action MUST be append_checklist or update_checklist
   ↪  if you found extractable values in the text just read.

  **After each append_checklist or update_checklist action:**
  - Verify whether all extractable values from the preceding text
   ↪  were included.
  - If you notice missed values, immediately append them as the
   ↪  next action before continuing.
```

Figure 35

System Prompt used in GAVEL-AGENT (Part 2/3)

```
# Document Reading Efficiency
- **NEVER** reread fully visited documents (marked with  Fully
↪  Visited).
- **NEVER** reread token ranges already viewed (shown as "Viewed
↪  tokens: X-Y").
- When reading partially visited documents (marked with Partially
↪  Visited), read ONLY unviewed token ranges.
- Check the "Viewed tokens" list before calling read_document to
↪  avoid redundant reads.

# Write Semantics
- **Any checklist item can have multiple values**; the
↪  `extracted` field is always a list.
- **append_checklist**: add new entries; **Do not** set Not
↪  Applicable via `append_checklist`.
- **update_checklist**: replace the entire `extracted` list; use
↪  for single-valued items, complete/authoritative sets,
↪  corrections, or to set "Not Applicable".

# Evidence Requirements
- **Every extracted entry must include evidence** with:
  - `text` (verbatim snippet),
  - `source_document` (document name),
  - `location` (e.g., page, section, docket entry; include token
  ↪  offsets if available).

# Not Applicable Encoding
- Represent Not Applicable as a **single extracted entry** for
↪  that key, set **via `update_checklist`**:
  - `value`: **"Not Applicable"** (exact string; case-sensitive)
  - `evidence`: required (explicit text or a dispositive posture
  ↪  supporting Not Applicable)
- A key is treated as **Not Applicable** only if its `extracted`
↪  list contains **exactly one** entry whose `value` is "Not
↪  Applicable".
- Do **not** mark Not Applicable solely because you failed to
↪  find a value; require explicit text or logically dispositive
↪  evidence (e.g., dismissal with prejudice -> no
↪  settlement/decree; "no class certification sought" -> class
↪  action items Not Applicable).
- If later evidence shows the item **does** have real values, use
↪  `update_checklist` to replace the Not Applicable entry with
↪  the confirmed entries.

# Stop Criteria
- Stop only when every checklist key is either:
  * Complete: all relevant values present in the corpus for that
  ↪  key have been extracted, each with evidence.
  * Not Applicable: represented as a single extracted entry with
  ↪  value "Not Applicable" and supporting evidence.
- Before stopping, verify state with `get_checklist` (in a prior
↪  turn if needed) and, if consolidation is required, issue one
↪  final `update_checklist` (in a prior turn) to replace any
↪  incrementally built keys with their curated final lists. Then
↪  return the stop decision.
```

Figure 36

System Prompt used in GAVEL-AGENT (Part 3/3)

~~~
{{TOOL_DESCRIPTIONS}}

# Response Format
- On each assistant turn, do exactly **one** of:
  1) **Issue one function call**, or
  2) **Stop** if all applicable checklist items are fully
   ↪  extracted and any non-applicable items are marked.
- When stopping, return **only** this JSON (no extra text):
```json
{
  "decision": "stop",
  "reason": "<brief justification>"
}
~~~

Figure 37