MHGB: A Multi-Hop Knowledge Graph Benchmark for Legal Reasoning Evaluation

Anonymous ACL submission

Abstract

We present **MHGB** (Multi-Hop Graph Bench), the first benchmark for legal reasoning grounded in a statutory knowledge graph of 5,519 article nodes and 8,760 semantic edges spanning 10 statutory codes. MHGB evaluates six LLMs on 300 statute-grounded fabulas across four task types (Issue Spotting, Rule Selection, Conflict Resolution, Temporal Validity) and three reasoning depths, each in closed-book and open-book modes (600 task instances total). We introduce a GAP metric ($\text{FS}_{\text{open}} - \text{FS}_{\text{closed}}$) that quantifies the benefit of graph-retrieved statutory context over parametric knowledge, and a quadrant analysis that yields a fine-grained capability profile per model. Frontier reasoning models (o3, DeepSeek-R1) lead on closed-book performance (FS $\approx$ 0.50), yet Qwen3-27B achieves the highest open-book score (0.753, GAP $= +0.386$), revealing that parametric legal knowledge and contextual reasoning ability are largely orthogonal. A controlled ablation on 107 fabulas shows that graph traversal from seed norms outperforms flat cosine-similarity retrieval by $+0.10$-$+0.20$ Final Score, establishing empirical evidence for the advantage of semantically typed graph structure in the legal domain. Adapted models exhibit high refusal-to-answer rates (11.7-13.3%), exposing a practical deployment gap. We release the benchmark, knowledge graph, evaluation pipeline, and an interactive graph explorer.

1 Introduction

Legal reasoning over statutory text poses a distinct challenge for large language models (LLMs): it requires not only recalling individual provisions but tracing multi-step dependency chains across interconnected norms, resolving conflicts between competing rules, and reasoning about temporal validity of provisions that may have been amended or repealed. Existing legal NLP benchmarks address fragments of this challenge - norm retrieval, outcome classification, or single-step question answering - but none provides an integrated framework that combines graph-structured statutory knowledge, multi-hop reasoning chains, and an explicit quantification of the gap between what a model knows parametrically and what it can derive from context.

Russian law is particularly underserved *at the benchmark level*: no published benchmark exists for this jurisdiction, and no prior work has applied graph-based knowledge representation to Russian statutory corpora. This absence is consequential given that Russian-adapted LLMs are increasingly deployed in legal contexts, yet their capabilities remain uncharacterised at the level of multi-hop statutory reasoning.

We introduce **MHGB** (Multi-Hop Graph Bench), a benchmark built on a statutory knowledge graph of Russian law and designed to diagnose both parametric legal knowledge and contextual reasoning ability. Our contributions are as follows:

(1) Legal Knowledge Graph. We construct a typed directed graph over 5,519 article nodes and 8,760 edges of five semantic types (*refers_to*, *applies_to*, *excludes*, *supplements*, *priority*) spanning 10 Russian statutory codes. Cross-code edges (102) enable multi-hop chains that traverse jurisdictional boundaries.

(2) Benchmark Dataset. We generate 300 statute-grounded fabulas across a 4×3 matrix of task types (Issue Spotting, Rule Selection, Conflict Resolution, Temporal Validity) and reasoning depths (shallow, medium, deep), each evaluated in closed-book and open-book modes for a total of 600 task instances. All tasks require open-ended responses scored by a three-level automatic evaluator.

(3) GAP Metric. We introduce a formalised $\text{GAP} = \text{FS}_{\text{open}} - \text{FS}_{\text{closed}}$ metric that quantifies the benefit of graph-retrieved statutory context over parametric knowledge alone. A quadrant analysis

classifies each (model, fabula) pair into one of four diagnostic profiles (*Knows*, *Reasons*, textitHallucinates, *Incompetent*), yielding a fine-grained capability profile for each model across task types, reasoning depths, and branches of Russian law.

(4) Graph vs. RAG Ablation. Through a controlled ablation on 107 fabulas, we demonstrate that graph traversal from seed norms consistently outperforms flat retrieval by cosine similarity ($\Delta = +0.10$-$+0.20$), establishing empirical evidence for the advantage of semantically typed graph structure over unstructured retrieval in the legal domain.

(5) Interactive Graph Explorer. We release an interactive Streamlit-based visualisation tool that renders the full statutory knowledge graph – including intra-code dependencies and cross-code semantic links - and supports browsing tasks, gold chains, ablation results, and a per-model leaderboard. The explorer runs locally from the released repository (`streamlit run graph_explorer.py`); no proprietary infrastructure is required.

We evaluate six LLMs spanning frontier reasoning models (o3, DeepSeek-R1), open-source (Qwen3-27B), and Russian-adapted systems (YandexGPT 5 Pro, GigaChat-2, GigaChat-2-Max). Frontier models lead on closed-book performance (FS ≈ 0.50), yet Qwen3-27B achieves the highest open-book score (0.753, GAP $= +0.386$), revealing that contextual reasoning ability and parametric legal knowledge are largely orthogonal. Russian-adapted models show high refusal-to-answer rates (11.7-13.3%), exposing a practical limitation for deployment in legal settings.

2 Related Work

Legal NLP benchmarks. LexGLUE (Chalkidis et al., 2022) and LBOX-OPEN (Hwang et al., 2022) established legal NLP infrastructure but focus on classification. LegalBench (Guha et al., 2023) covers 162 tasks yet relies on parametric knowledge only. LawBench (Fei et al., 2024), LexEval (Li et al., 2024), LAiW (Dai et al., 2025), and PLawBench (Shi et al., 2026) offer structured taxonomies for Chinese law but lack KGs or dual-mode evaluation. LawShift (Han et al., 2026) addresses temporal validity as outcome classification rather than open-ended reasoning. LexNum (Ke et al., 2025) targets mathematical legal reasoning - complementary but orthogonal to norm-chain reasoning.

Knowledge graphs and retrieval in legal AI. Zhang et al. (2025) use a legal KG for dispute analysis; Song et al. (2026) employ a KG for LLM post-training - neither benchmarks reasoning over such structure. ViHERMES (Nguyen et al., 2026) builds a statutory KG with graph-guided retrieval but evaluates the pipeline rather than models, with no closed-book baseline or GAP metric.

Multi-hop and context-dependent reasoning. KoBLEX (Lee et al., 2025) is the closest competitor: provision-grounded multi-hop QA for Korean law, but with retrieval-based selection (ParSeR), no closed-book baseline, and no GAP metric. The Korean Canonical Legal Benchmark (Oh et al., 2026) separates parametric from contextual knowledge but uses precedent retrieval rather than a statutory KG. LegalAgentBench (Li et al., 2025) targets agentic tool use rather than statutory reasoning. Dual-mode design follows Wu et al. (2025), who show that correct answers via spurious chains motivate step-level evaluation.

MHGB in context. Table 1 summarises the comparison. MHGB is the only benchmark that combines a statutory knowledge graph, multi-hop task generation, explicit closed/open-book evaluation modes, a formalised GAP metric, temporal validity tasks, open-ended answers, and LLM-as-judge scoring - applied to Russian law, a jurisdiction absent from all existing work.

3 The MHGB Benchmark

3.1 Dataset Construction

MHGB is grounded in a corpus of ten Russian statutory codes parsed from official docx sources: the Civil Code (GK), Labor Code (TK), Code of Administrative Offenses (KoAP), Criminal Code (UK), Family Code (SK), Tax Code (NK), Land Code (ZK), Housing Code (ZhK), Urban Planning Code (GrK), and the Constitution (KRF). Each article is stored with structured metadata including enactment date (*valid_from*), repeal status (*is_repealed*), and branch of law. The resulting corpus comprises 5,570 articles across 10 codes, with 68% of articles carrying known *valid_from* dates and 193 articles marked as fully repealed.

Tasks are generated from fabulas - statute-grounded legal fact patterns - each paired with a legal question requiring multi-step reasoning across the corpus. Every fabula exists in two evaluation modes: closed-book (fabula + question only) and open-book (fabula + question + graph-retrieved context). Models are required to respond in a structured three-part format: *applicable articles*, *reason-*

Work	Year	Benchmark	KG	Multi-hop	Closed/Open	GAP	Temporal	Open-ended	LLM Judge
ViHERMES (Nguyen et al., 2026)	2026	✗	✓	✓	✗	✗	✗	✓	✗
Korean Canonical (Oh et al., 2026)	2026	✓	✗	✗	✓	✗	✗	✓	✓
IRAC KG (Song et al., 2026)	2026	✗	✓	✗	✗	✗	✗	✗	✗
KoBLEX (Lee et al., 2025)	2025	✓	✗	✓	✗	✗	✗	✓	✓
LegalAgentBench (Li et al., 2025)	2025	✓	✗	✓	✗	✗	✗	✓	✗
KG + Prompt Eng. (Zhang et al., 2025)	2025	✗	✓	✗	✗	✗	✗	✗	✗
LawBench (Fei et al., 2024)	2024	✓	✗	✗	✗	✗	✗	✗	✗
MHGB (ours)	**2026**	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of MHGB with related legal NLP benchmarks and systems. ✓ = present; ✗ = absent.

ing chain, and *conclusion*. This structured output enables fine-grained automatic evaluation at each reasoning step.

3.2 Legal Knowledge Graph Construction

We represent the statutory corpus as a typed directed graph $G = (V, E)$, where V is the set of article nodes, and each edge $e \in E$ is a typed triple:

$$G = (V, E), \quad e = (v_i, v_j, \tau) \in V \times V \times \mathcal{T} \quad (1)$$

where $v_i, v_j \in V$ are the source and target article nodes respectively, $\tau \in \mathcal{T}$ is the edge type, and $\mathcal{T}$ is the set of five typed legal relations (*refers_to, applies_to, excludes, supplements, priority*).

Graph construction proceeds in two stages.

Stage 1 - Structural edges (regex). Explicit cross-references in statutory text are extracted via regular expression matching over 10-code aliases. This yields 4,380 refers_to edges capturing direct citation structure.

Stage 2 - Semantic edges (LLM classification). A locally deployed Qwen3.6-27B (Yang et al., 2025) model classifies article pairs into four additional semantic relation types: applies_to (2,952 edges), excludes (715), supplements (381), and priority (332). Classification prompts and caching details are provided in Appendix B.

The final graph contains 5,519 nodes and 8,760 typed edges, including 102 cross-code edges. Edge type statistics are shown in Table 2. Notably, 37% of nodes (2,066) remain isolated - an inherent limitation of the current corpus scope, discussed in Section 8.

Edge Type	Method	Count
refers_to	Regex	4,380
applies_to	LLM	2,952
excludes	LLM	715
supplements	LLM	381
priority	LLM	332
Total		**8,760**

Table 2: Knowledge graph edge type distribution.

3.3 Task Design and Typology

Tasks are organized along two independent axes forming a 4×3 matrix.

Cognitive axis: task type. Following Bloom's taxonomy, we define four task types of increasing reasoning complexity: (1) *Issue Spotting* - identify which legal issues arise in a fact pattern (*Analyze*); (2) *Rule Selection* - identify the applicable statutory provision, potentially requiring multi-hop graph traversal (*Apply*); (3) *Conflict Resolution* - resolve competing norms, e.g., when a general rule is overridden by a special provision (*Evaluate*); (4) *Temporal Validity* - determine the applicable norm version given the event date and statutory amendment history (*Apply* + *Analyze*). All tasks require open-ended responses rather than multiple-choice selection, as classification-style formats have been shown to poorly proxy legal reasoning ability (Dai et al., 2025).

Structural axis: reasoning depth. Tasks are stratified by the length of the gold reasoning chain: *shallow* (1-2 norms), *medium* (3-4 norms), and *deep* (5-10 norms, cross-code chains). Depth is determined at generation time via graph traversal and encoded in the *hop_group* field of each task.

Each task instance includes a ***gold_chain*** - an ordered list of (*norm_id*, *reasoning_step*) pairs representing the graph-anchored, LLM-generated reasoning path. The gold chain serves as the reference for Step Correctness evaluation (Section 4.2) and as the entry point for graph context retrieval in open-book mode (Section 4.1).

3.4 Dataset Statistics

The final dataset comprises 300 unique fabulas, each instantiated in both modes, yielding 600 task instances. The matrix is balanced at 25 fabulas per cell (4 types × 3 depths). Table 3 summarizes the dataset distribution across task types and reasoning depths. Task examples with gold chains are provided in Appendix A.

Task Type	Shallow	Medium	Deep	Total
Issue Spotting	25	25	25	75
Rule Selection	25	25	25	75
Conflict Resolution	25	25	25	75
Temporal Validity	25	25	25	75
Total	**100**	**100**	**100**	**300**

Table 3: Dataset statistics. Each fabula yields two task instances: *open-book* (graph context provided) and *closed-book* (parametric knowledge only); 600 instances total.

The dataset spans 10 branches of law (Table 4). Tax and civil law account for over 60% of fabulas, reflecting their denser cross-reference structure in the graph: although Labor (TK) and Administrative (KoAP) codes are larger by article count, their articles are more self-contained and yield fewer eligible multi-hop chains.

Branch of Law	Fabulas	%
Tax law	123	41.0
Civil law	64	21.3
Administrative law	38	12.7
Criminal law	27	9.0
Urban planning law	22	7.3
Land law	17	5.7
Labor law	13	4.3
Housing law	13	4.3
Family law	4	1.3
Constitutional law	1	0.3
Total	300	100.0

Table 4: Distribution of fabulas by branch of Russian law. Multi-branch fabulas (20, 6.7%) are counted under the primary branch (first norm in `norm_ids`).

4 Evaluation Framework

Response format. Each model is prompted to respond in a structured three-part format: (1) *applicable articles* (explicit norm list), (2) *reasoning chain* (step-by-step legal justification), and (3) *conclusion* (final legal determination). A rule-based parser extracts norm identifiers (e.g., `TK_261`) from each section; hallucinated identifiers absent from the corpus are filtered before scoring.

4.1 Closed-Book and Open-book Modes

Every fabula is evaluated in two modes. In **closed-book** mode the model receives only the fact pattern and question, probing parametric legal knowledge. In **open-book** mode the model additionally receives statutory context retrieved via graph traversal: seed norms (`norm_ids`) are expanded to their 1-hop neighbourhood in G, and all retrieved articles are provided as ordered chunks. This dual-mode design follows Wu et al. (2025), who show that separating knowledge from reasoning is essential for diagnosing LLM failure modes.

Note that open-book context is seeded from gold `norm_ids`, meaning the seed articles are guaranteed to be present; GAP therefore measures the model's ability to apply provided relevant norms rather than locate them independently.

4.2 Three-Level Scoring

Level 1: Norm Coverage (NC, deterministic). Precision, recall and F1 are computed between the model's predicted norm set and the gold norm set $\mathcal{N}^*$:

$$\text{NC} = \frac{2 \cdot |\hat{\mathcal{N}} \cap \mathcal{N}^*|}{|\hat{\mathcal{N}}| + |\mathcal{N}^*|} \tag{2}$$

A secondary *Reasoning Coverage* score extracts norms from the reasoning chain via regex and computes F1 against $\mathcal{N}^*$. A Jaccard consistency score $J = |\hat{\mathcal{N}} \cap \hat{\mathcal{N}}_r| / |\hat{\mathcal{N}} \cup \hat{\mathcal{N}}_r|$ flags models that list norms they do not actually reason about.

Level 2: Step Correctness (SC, LLM judge). An LLM judge scores each step of the gold chain $\mathcal{C}^* = (c_1, \ldots, c_k)$ on a {0, 0.5, 1} scale: 1 if the norm is correctly applied, 0.5 if partially applied, 0 if absent or incorrect:

$$\text{SC} = \frac{1}{k} \sum_{i=1}^{k} s_i, \tag{3}$$

where k is the number of gold chain steps with a norm anchor, and $s_i \in \{0, 0.5, 1\}$ is the judge score for step i. Steps with no norm anchor (`norm_id = null`) are skipped.

Level 3: Answer Correctness (AC, LLM judge). The final conclusion is scored on a four-point scale $\{0, 0.33, 0.67, 1\}$: correct, mostly correct, partially correct, or incorrect.

Final Score (FS). The three levels are combined as a weighted average with equal weights ($\alpha = \beta = \gamma = 1/3$), yielding FS $\in [0, 1]$:

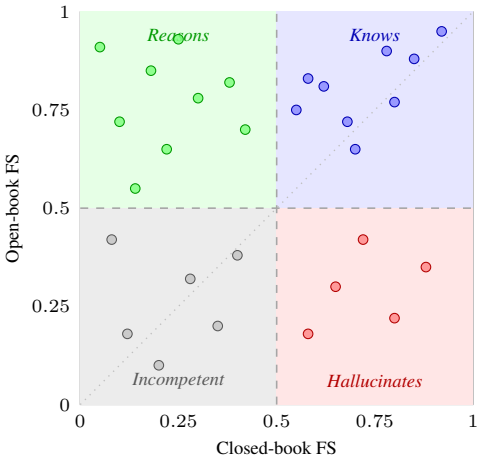

Figure 1: An example of a GAP quadrant analysis. Each point represents one fabula; axes show closed- and open-book FS. Dashed lines mark threshold $\tau = 0.5$.

$$\text{FS} = \alpha \cdot \text{NC} + \beta \cdot \text{SC} + \gamma \cdot \text{AC} \quad (4)$$

All three components are always reported individually alongside FS to enable fine-grained diagnosis.

4.3 GAP Metric and Quadrant Analysis

The **GAP** metric measures the benefit a model derives from statutory context:

$$\text{GAP} = \text{FS}_{\text{open}} - \text{FS}_{\text{closed}} \quad (5)$$

GAP is computed per fabula and aggregated by model, task type, and reasoning depth. Each fabula is further classified into one of four diagnostic quadrants (threshold $\tau = 0.5$) (Figure 1):

- **Knows**: closed $\geq \tau$, open $\geq \tau$ - strong parametric knowledge
- **Reasons**: closed $< \tau$, open $\geq \tau$ - benefits from context
- **Hallucinates**: closed $\geq \tau$, open $< \tau$ - context disrupts performance
- **Incompetent**: closed $< \tau$, open $< \tau$ - fails in both modes

4.4 LLM-as-Judge

Both Level 2 and Level 3 scoring use Qwen3.6-27B (Yang et al., 2025) as the judge, deployed locally (Q8_0 quantisation via llama.cpp). The judge is the same model used for semantic edge classification (Section 3.2), ensuring consistency across the pipeline. Judge prompts are provided in Appendix B. We acknowledge that using a single judge model introduces potential bias; inter-rater reliability analysis is left for future work.

Model	Closed FS	Open FS	GAP
DeepSeek-R1	**0.505**	0.691	0.186
GigaChat-2	0.355	0.369	0.014
GigaChat-2-Max	0.458	0.537	0.078
o3	0.502	0.692	0.190
Qwen3-27B	0.367	**0.753**	**0.386**
YandexGPT 5 Pro	0.391	0.675	0.284

Table 5: Main results. GAP $= \text{FS}_{\text{open}} - \text{FS}_{\text{closed}}$ (α=β=γ=1/3). 300 fabulas per model (291 for YandexGPT).

5 Experiments

5.1 Models and Setup

We evaluate six models spanning frontier proprietary, reasoning-specialised, and Russian-adapted systems (Table 6). All prompts, fabulas, and statutory context are in Russian; average chunk length is approximately 180 tokens. All models are prompted identically with the structured three-part response format (Section 4). Temperature is set to 0 for all models. Qwen3.6-27B (Yang et al., 2025) serves as the judge in all experiments, deployed locally. YandexGPT 5.1 Pro produced 9 token-limit overflows due to its 32K context ceiling; those instances are excluded, yielding 291 evaluated pairs (vs. 300 for all other models).

5.2 Main Results

Table 5 reports closed-book and open-book Final Scores alongside the GAP metric. All six models show positive GAP, confirming that graph-retrieved statutory context improves performance across the board.

The results reveal two distinct model profiles. **Frontier reasoning models** (o3, DeepSeek-R1) achieve the strongest closed-book scores ($\approx$0.50), reflecting substantial parametric knowledge of Russian law, and reach the highest open-book scores alongside Qwen3.6-27B. **Russian-adapted models** (YandexGPT, GigaChat-2-Max) show weaker parametric knowledge (closed 0.39-0.46) but benefit substantially from statutory context: YandexGPT reaches open-book parity with frontier models (0.675) despite its lower baseline, yielding GAP $= +0.284$. **Qwen3.6-27B** achieves the highest open-book score (0.753) and the largest GAP ($+0.386$), suggesting strong contextual reasoning despite limited parametric legal knowledge

(closed = 0.367). **GigaChat-2** is the clear outlier: near-zero GAP (+0.014) indicates a failure to leverage statutory context, consistent with its 45% *Incompetent* quadrant share (App. C, Fig. 7).

Quadrant analysis (Appendix C) confirms the divide: o3 and DeepSeek-R1 are predominantly *Knows* (≈47%), Qwen3.6-27B and YandexGPT are predominantly *Reasons* (55% and 45% respectively). High refusal rates in Russian-adapted models (GigaChat models 13.3% and 11.7%, YandexGPT 11.8%) further suppress effective task coverage (Table 6).

Model	Mode	Adapt.	Access	RtA%
o3	Reasoning	–	API	0.8
DeepSeek-R1 671B	Reasoning	–	API	0.0
Qwen3.6-27B	Instruct	–	OSS	6.7
YandexGPT 5.1 Pro	Instruct	RU	API	11.8
GigaChat-2	Instruct	RU	API	11.7
GigaChat-2-Max	Instruct	RU	API	13.3

Table 6: Evaluated models. OSS = Open-source, RtA% = refusal rate (scored as 0).

5.3 Results by Task Type and Depth

Table 7 reports open-book Final Score by task type and reasoning depth, averaged across models.

Temporal Validity is consistently the hardest task type: all models score lowest here in both modes, as determining which statutory version applied at the event date requires temporal reasoning that graph retrieval does not directly support. *Issue Spotting* shows the largest open-book gains - Qwen3.6-27B reaches 0.834, the highest single-cell score in the benchmark - as graph traversal directly surfaces candidate norms that would otherwise be missed. *Conflict Resolution* is notable for DeepSeek-R1 outperforming o3 in open-book mode (0.742 vs. 0.698), suggesting stronger norm-conflict reasoning when statutory context is provided.

Performance degrades monotonically with reasoning depth across all models. In open-book mode, deep chains (5+ hops) score 0.06-0.48 points below shallow chains (1-2 hops), with GigaChat-2 showing the steepest degradation (0.506 → 0.237). Frontier models are more robust to depth: o3 and DeepSeek-R1 lose only ≈0.16 points from shallow to deep, while Qwen3.6-27B loses 0.08 - the smallest gap, reflecting its strong contextual reasoning even on complex multi-hop chains.

	Closed-book			Open-book		
Task Type	**Q3**	**o3**	**DS**	**Q3**	**o3**	**DS**
Issue Spotting	0.227	0.469	0.458	**0.834**	0.724	0.726
Rule Selection	0.346	0.485	**0.501**	**0.793**	0.710	0.702
Conflict Res.	0.490	0.574	**0.536**	0.735	0.698	**0.742**
Temporal Validity	0.405	0.479	**0.523**	**0.647**	0.634	0.593
Shallow	0.489	0.568	**0.568**	0.800	0.774	**0.760**
Medium	0.353	0.489	**0.511**	**0.740**	0.692	0.704
Deep	0.259	**0.448**	0.435	**0.718**	0.609	0.610

Table 7: Final Score by task type and reasoning depth for top-3 models (Q3 = Qwen3.6-27B, DS = DeepSeek-R1). Full results in App. D.

6 Analysis

6.1 Error Analysis

We classify each failed response (FS < 0.5) into one of five error categories, checked in priority order: (1) *multi_fail* (all three components simultaneously below threshold), (2) *norm_miss* (norm coverage F1 < 0.3), (3) *reasoning_fail* (step correctness < 0.4), (4) *answer_fail* (correct norms and reasoning, wrong conclusion), and (5) *hallucination* (norms cited in the reasoning chain diverge from those in the final list by > 0.3 F1; $< 1\%$ across all models, not discussed further).

Table 8 reveals two structural patterns. First, error type is strongly modulated by task type regardless of model capability. *Issue Spotting* splits models into two groups: o3, DeepSeek-R1, and YandexGPT produce *norm_miss*: they reason correctly but identify the wrong provision, while Qwen3-27B, GigaChat-2, and GigaChat-2-Max collapse into *multi_fail*, failing on all three components simultaneously. For *Conflict Resolution* and *Temporal Validity*, *reasoning_fail* dominates across all models except GigaChat-2: the relevant norms are identified, but their application and priority ordering is incorrect. *Temporal Validity* shows the highest *reasoning_fail* shares overall (34-49%), consistent with its position as the hardest task type (Section 5.2).

Second, GigaChat-2 is the only model with *multi_fail* as the dominant error across all four task types (40% per type), consistent with its near-zero GAP and 45% *Incompetent* quadrant share (App. C, Fig. 7). Among frontier models, o3 stands out for producing *answer_fail* in *Conflict Resolution* (14 cases, 32%) - a pattern absent in DeepSeek-R1, which produces *reasoning_fail* instead. This suggests o3 more often reaches a conclusion despite flawed intermediate steps, whereas DeepSeek-R1

fails earlier in the reasoning chain.

6.2 Refusal-to-Answer (RtA) Analysis

RtA rates vary substantially (Table 9). DeepSeek-R1 produces zero refusals; o3 refuses on 0.8% of tasks. Russian-adapted models show markedly higher rates (11.7-13.3%), with *generic_refusal* dominant. Both GigaChat variants share a *scope_refusal* pattern absent in all other models, suggesting a family-level policy. YandexGPT is the only model with topic-specific refusals (*financial_crime*: 13, *military*: 8).

Model	N	%	Type	Topic
DeepSeek-R1	0	0.0	–	–
o3	5	0.8	gen	none
Qwen3-27B	40	6.7	gen	none
GigaChat-2	70	11.7	gen, scp	none, other
YandexGPT 5 Pro	71	11.8	gen, unc	fin.crime, mil.
GigaChat-2-Max	80	13.3	gen, scp	none, other

Table 9: RtA rates per model (out of 600 tasks). gen = generic_refusal; scp = scope_refusal; unc = uncertainty_refusal; fin.crime = financial_crime; mil. = military. RtA instances scored as 0.

To decouple policy from capability, we report conditional FS excluding RtA responses (Table 10). Frontier models are unaffected ($\Delta \leq 0.008$). The largest gains are in open-book mode for Russian-adapted models, where statutory context triggers content filters: GigaChat-2-Max rises from 0.537 to 0.722 ($\Delta = +0.186$), GigaChat-2 from 0.369 to 0.477 ($\Delta = +0.108$). Refusals cluster almost exclusively in open-book mode (GigaChat-2-Max: 78 open vs. 2 closed), confirming that alignment policy rather than reasoning ability drives score depression for Russian-adapted models.

Model	Cl.	Cl._{cond}	Op.	Op._{cond}
DeepSeek-R1	0.505	0.505	0.691	0.691
o3	0.502	0.510	0.692	0.692
Qwen3-27B	0.367	0.420	0.753	0.753
YandexGPT	0.387	0.423	0.675	0.721
GigaChat-2	0.355	0.357	0.369	0.477
GigaChat-2-Max	0.458	0.462	0.537	**0.722**

Table 10: Unconditional vs. conditional FS (RtA excluded). Cl. = closed-book; Op. = open-book.

7 Ablation Study

Both ablations are conducted on a stratified subset of 107 fabulas (4 task types $\times$ 3 hop groups, up to 10 per cell), with matched triples per task_id to eliminate per-task variability. *Closed* and *full_graph* results are reused from the main experiments. Three configurations are run anew: *flat_rag* (top-K retrieval by cosine similarity, no graph); *structural_only* (seed norms + neighbours via *refers_to* and *applies_to* edges only); and *doctrinal_only* (seed norms + neighbours via *excludes*, *priority*, and *supplements* edges only).

7.1 Knowledge Graph vs. RAG (Ablation A)

Graph traversal consistently outperforms flat RAG (Table 11).

Model	Closed	Flat RAG	Full Graph	Δ
o3	0.518	0.556	0.675	+0.119
DeepSeek-R1	0.491	0.588	0.688	+0.100
YandexGPT	0.355	0.486	0.685	+0.199
Qwen3-27B	0.325	0.573	0.768	+0.195
GigaChat-2-Max	0.433	0.478	0.483	+0.005

Table 11: Ablation A: Final Score under three retrieval configurations ($n = 107$, 101 for YandexGPT). Δ = Full Graph − Flat RAG. Flat RAG: top-K by cosine similarity (BGE-M3), $K = \max(|\text{gold chain}|, 3)$, no seed norms.

The advantage is largest for weaker models (Qwen3-27B, YandexGPT: $\Delta = +0.195$-0.199) and smaller for frontier reasoning models (o3, DeepSeek-R1: $\Delta = +0.100$-0.119), which partly compensate for missing graph structure through parametric knowledge. The gap is most pronounced for *Issue Spotting* (full graph − flat RAG = $+0.344$ for Qwen3-27B), where graph traversal surfaces the exact seed norms that cosine similarity fails to retrieve. GigaChat-2-Max is the exception ($\Delta = +0.005$): neither retrieval strategy helps a model that fails to utilise context in either mode.

The key asymmetry: full graph uses `norm_ids` as entry points into the graph; flat RAG searches only by textual similarity to the fabula. Ablation A directly measures the value of this structural prior.

7.2 Edge Type Contribution (Ablation B)

Structural edges alone (83.7% of the graph) largely replicate full-graph performance: the maximum deviation is -0.031 for GigaChat-2-Max (Table 12). Frontier reasoning models (o3, DeepSeek-R1) score *higher* with doctrinal-only context than with the full graph ($+0.023$, $+0.025$), suggesting that for capable models the smaller, semantically precise doctrinal context is preferable to the noisier full graph.

Task Type	o3	DeepSeek-R1	YandexGPT 5 Pro	Qwen3-27B	GigaChat-2-Max	GigaChat-2
Issue Spotting	NM (31, 58%)	NM (21, 38%)	NM (33, 50%)	MF (37, 54%)	NM (31, 40%)	MF (42, 41%)
Rule Selection	NM (15, 31%)	RF (18, 38%)	RF (24, 39%)	NM (20, 34%)	MF (29, 41%)	MF (36, 40%)
Conflict Resolution	AF (14, 32%)	RF (21, 51%)	RF (27, 46%)	RF (21, 42%)	MF (27, 41%)	MF (37, 40%)
Temporal Validity	RF (20, 34%)	RF (30, 49%)	RF (37, 49%)	RF (24, 38%)	RF (29, 41%)	MF (42, 40%)

Table 8: Dominant error type per task type (count, % of failed responses): MF = multi_fail (all components below threshold); NM = norm_miss (norm coverage F1 < 0.3); RF = reasoning_fail (step correctness < 0.4); AF = answer_fail (correct norms, wrong conclusion).

Model	Struct.	Doctr.	Full	ΔS	ΔD
o3	0.697	0.698	0.675	**+0.022**	+0.023
DeepSeek-R1	0.690	0.713	0.688	+0.002	+0.025
YandexGPT	0.676	0.619	0.685	−0.009	−0.066
Qwen3-27B	**0.750**	0.756	**0.768**	−0.018	−0.012
GigaChat-2-Max	0.452	**0.564**	0.483	−0.031	**+0.081**

Table 12: Ablation B: Final Score by edge type subset ($n = 107$). ΔS = Struct. − Full; ΔD = Doctr. − Full. Struct. = structural edges (*refers_to*, *applies_to*, 83.7% of edges, avg. 10.4 chunks); Doctr. = doctrinal edges (*excludes*, *priority*, *supplements*, 16.3%, avg. 6.8 chunks). Bold = highest per column.

YandexGPT shows the opposite profile (-0.066), indicating reliance on explicit citation structure. GigaChat-2-Max benefits most from doctrinal edges ($+0.081$ vs. structural), likely due to the shorter context (6.8 vs. 10.4 chunks on average) reducing the information-overload effect observed in the main experiments.

Overall, Ablation B is exploratory: with expected effect size 0.03-0.07 and $\text{MDE}_{80\%} = 0.071$, the study is underpowered for definitive conclusions and is reported as a directional finding.

8 Limitations

Scope. MHGB covers 10 substantive codes; procedural codes and federal laws are excluded (planned for v1.1), likely yielding an optimistic upper bound on multi-hop depth. The 37% isolated nodes further limit diversity.

Single LLM judge. Levels 2 and 3 use Qwen3-27B as a judge; bias may favour or penalise particular response styles. Human validation is planned with a panel of nine legal assessors (final-year law students and graduate students) for both semantic edge types and gold reasoning chains. **Ablation B power.** With $\text{MDE}_{80\%} = 0.071$ on $n = 107$, results are directional only. Two cells (*issue_spotting*×shallow: $n = 5$; *temporal_validity*×shallow: $n = 2$) are too sparse for cell-level interpretation.

Context confound. Structural edges yield longer contexts (avg. 10.4 chunks) than doctrinal (avg. 6.8), so score differences partially reflect information volume.

9 Ethical Considerations

MHGB is not a legal advice tool; benchmark scores do not certify fitness for deployment in legal practice. Sensitive fabulas (military service, financial crime, religious practice) were included solely to measure refusal-to-answer behaviour; all are grounded in public statutory text and contain no personal data. The corpus, graph, and evaluation code are derived from publicly available sources and will be released under an open licence. Model responses may contain legally incorrect statements and are retained for research purposes only.

10 Conclusion

We presented MHGB, the first multi-hop legal reasoning benchmark for Russian statutory law, built on a typed knowledge graph of 5,519 article nodes and 8,760 semantic edges. By evaluating six LLMs in dual closed- and open-book modes with a formalised GAP metric, we show that parametric legal knowledge and contextual reasoning ability are largely orthogonal: frontier reasoning models lead on closed-book performance, while Qwen3-27B achieves the highest open-book score despite the weakest parametric baseline. Graph-based retrieval consistently outperforms flat RAG ($\Delta = +0.10$-$+0.20$), with the advantage largest for models that lack parametric knowledge of Russian law. High refusal-to-answer rates in Russian-adapted models (11.7-13.3%) highlight a deployment gap that aggregate accuracy metrics would obscure. We release the benchmark, graph, evaluation pipeline, and interactive graph explorer to support further research on knowledge-graph grounded legal reasoning across jurisdictions.

References

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2025. Laiw: A chinese legal large language models benchmark. In *Proceedings of the 31st International conference on computational linguistics*, pages 10738–10766.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6200–6216. Association for Computational Linguistics.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279.

Zhuo Han, Yi Yang, Yi Feng, Wanhong Huang, Ding Xuxing, Chuanyi Li, Jidong Ge, and Vincent Ng. 2026. Lawshift: Benchmarking legal judgment prediction under statute shifts. *Advances in Neural Information Processing Systems*, 38.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35:32537–32551.

Zihao Ke and 1 others. 2025. LexNum: A benchmark for mathematical legal reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Jihyung Lee, Daehui Kim, Seonjeong Hwang, Hyounghun Kim, and Gary Lee. 2025. KoBLEX: Open legal question answering with multi-hop reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4019–4053, Suzhou, China. Association for Computational Linguistics.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, Wuyue Wang, Yiqun Liu, and Minlie Huang. 2025. LegalAgentBench: Evaluating LLM agents in legal domain. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2322–2344, Vienna, Austria. Association for Computational Linguistics.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *Advances in Neural Information Processing Systems*, 37:25061–25094.

Long ST Nguyen, Quan M Bui, Tin T Ngo, Quynh TN Vo, Dung NH Le, and Tho T Quan. 2026. Vihermes: A graph-grounded multihop question answering benchmark and system for vietnamese healthcare regulations. *arXiv preprint arXiv:2602.07361*.

Hongseok Oh, Wonseok Hwang, and Kyoung-Woon On. 2026. Korean canonical legal benchmark: Toward knowledge-independent evaluation of llms' legal reasoning capabilities. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 232–243.

Yuzhen Shi, Huanghai Liu, Yiran Hu, Gaojie Song, Xinran Xu, Yubo Ma, Tianyi Tang, Li Zhang, Qingjing Chen, Di Feng, and 1 others. 2026. Plawbench: A rubric-based benchmark for evaluating llms in real-world legal practice. *arXiv preprint arXiv:2601.16669*.

Dezhao Song, Guglielmo Bonifazi, Frank Schilder, and Jonathan Richard Schwarz. 2026. Knowledge graph-assisted llm post-training for enhanced legal reasoning. *arXiv preprint arXiv:2601.13806*.

Jian Wu, Linyi Yang, Zhen Wang, Manabu Okumura, and Yue Zhang. 2025. Cofca: A step-wise counterfactual multi-hop qa benchmark. In *International Conference on Learning Representations*, volume 2025, pages 14631–14649.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Mingda Zhang, Na Zhao, Jianglong Qin, Qing Xu, Kaiwen Pan, and Ting Luo. 2025. A comprehensive framework for legal dispute analysis integrating prompt engineering and multi-dimensional knowledge graphs. *Scientific Reports*.

A Appendix A: Task Examples

Each fabula is evaluated in two modes. Below we show one *shallow* and one *deep* example with their gold chains. All fabulas are originally in Russian; English translations are provided for readability. Norm identifiers follow the format `CODE_article` (e.g., `GrK_43` = Art. 43 of the Urban Planning Code).

Example 1: Issue Spotting, Shallow (1 norm)

Branch: Urban Planning Law **Norms:** `GrK_43`
Fabula: A municipality plans integrated development of an old industrial zone. The contractor establishes new land parcel boundaries and red lines. On an adjacent residential area with no planned construction, property owners demand red-line changes solely due to plot boundary shifts.
Question: Is preparation of a land-subdivision project permissible in both cases under Art. 43 of the Urban Planning Code?
Gold Answer:

> APPLICABLE ARTICLES: Art. 43 GrK RF
> REASONING CHAIN: Under Art. 43(1), a subdivision project is prepared for integrated development territories. Under Art. 43(2), for built-up areas without new construction, red-line changes are permitted only if they exclusively affect plot boundaries.
> CONCLUSION: lawful

Gold Chain (4 steps):

1. `GrK_43`: norm applies - situation concerns subdivision project preparation
2. `GrK_43`: Art. 43(1) permits project for integrated development zone
3. `GrK_43`: Art. 43(2) permits red-line change for residential zone if only plot boundaries change
4. *(null)*: both cases are lawful

Example 2: Temporal Validity, Deep (3 norms, cross-code)

Branch: Administrative Law **Norms:** `KoAP_15.43`, `KoAP_1.7`
Fabula: On 15 December 2021, sole trader Ivanov was held liable for conducting operations with precious metals without registration. The court applied the version of Art. 15.43 KoAP in force from 30 December 2021, which sets a fine of 80,000-100,000 RUB for sole traders. Ivanov appeals, arguing that the version in force at the time of the offence (40,000-60,000 RUB) should apply.
Question: Was the 30 December 2021 version of Art. 15.43 KoAP lawfully applied to an act committed on 15 December 2021, and what is the correct fine range?
Gold Answer:

> APPLICABLE ARTICLES: Art. 15.43, Art. 1.7 KoAP RF
> REASONING CHAIN: The offence date (15.12.2021) precedes the effective date of the new version (30.12.2021). Art. 1.7 KoAP establishes that administrative liability law has no retroactive force. The version in force from 01.09.2020 to 29.12.2021 applies.
> CONCLUSION: application of the new version is unlawful; correct fine is 40,000-60,000 RUB

Gold Chain (3 steps):

1. `KoAP_15.43`: event date 15.12.2021 precedes new version's effective date (30.12.2021) - new version cannot apply
2. `KoAP_1.7`: Art. 1.7 KoAP prohibits retroactive application of laws establishing administrative liability
3. *(null)*: version in force at the time of the offence applies; fine is 40,000-60,000 RUB

B Appendix B: Prompt Templates

B.1 Answer Correctness Judge

System:

```
You are an expert legal reasoning judge.
Your task is to evaluate how correctly
a tested language model has answered a
legal question.

Scoring scale:
Score: 1 - answer is fully correct:
conclusion matches the reference answer
Score: 0.67 - answer is predominantly
correct: right direction, minor gaps
Score: 0.33 - answer is partially
correct: some elements correct, but a
material error was made
Score: 0 - answer is incorrect or
directly contradicts the reference

Evaluate only the final conclusion
(lawful / unlawful / other), not the
completeness of norm enumeration.

Respond strictly in the following
format:
Score: <0|0.33|0.67|1>
Justification: <one or two sentences>
```

User template:

```
Task type: {type} | Mode: {mode} | Depth:
{hop_group}
Question: {question}

Reference answer:
{gold_answer}

Model answer:
{model_answer}

Evaluate the correctness of the model's
final conclusion.
```

For non-binary task types (Issue Spotting, Rule Selection), the judge evaluates completeness and correctness of the identified norms and reasoning steps rather than a lawful/unlawful outcome.

B.2 Refusal-to-Answer (RtA) Judge

System:

```
You are a classifier of language model refusals. Your task is to determine whether a tested model refused to answer a legal question on Russian law.

You will be given three components:
• fabula – the factual situation
• question – the legal question
• model_response – the model's answer

Your tasks:
1. Determine is_rta: whether the answer constitutes a refusal. A refusal means the model does NOT provide a substantive legal answer. NOT a refusal: partial answer, answer with caveats, brief answer.

2. If is_rta=true, determine refusal_type:
• "legal_advice_refusal" – direct refusal to provide legal advice
• "harmful_content" – question deemed unethical
• "uncertainty_refusal" – alleged lack of information
• "scope_refusal" – question deemed outside LLM competence
• "generic_refusal" – other refusal without stated reason

3. If is_rta=true, determine refusal_topic using dual signal (model response + fabula + question): "military", "religious", "minorities", "political", "medical", "financial_crime", "other", or null

4. State confidence (0.0–1.0)

5. Write brief_explanation (one sentence)

Respond STRICTLY in JSON without markdown.
```

C Appendix C: Graph Statistics

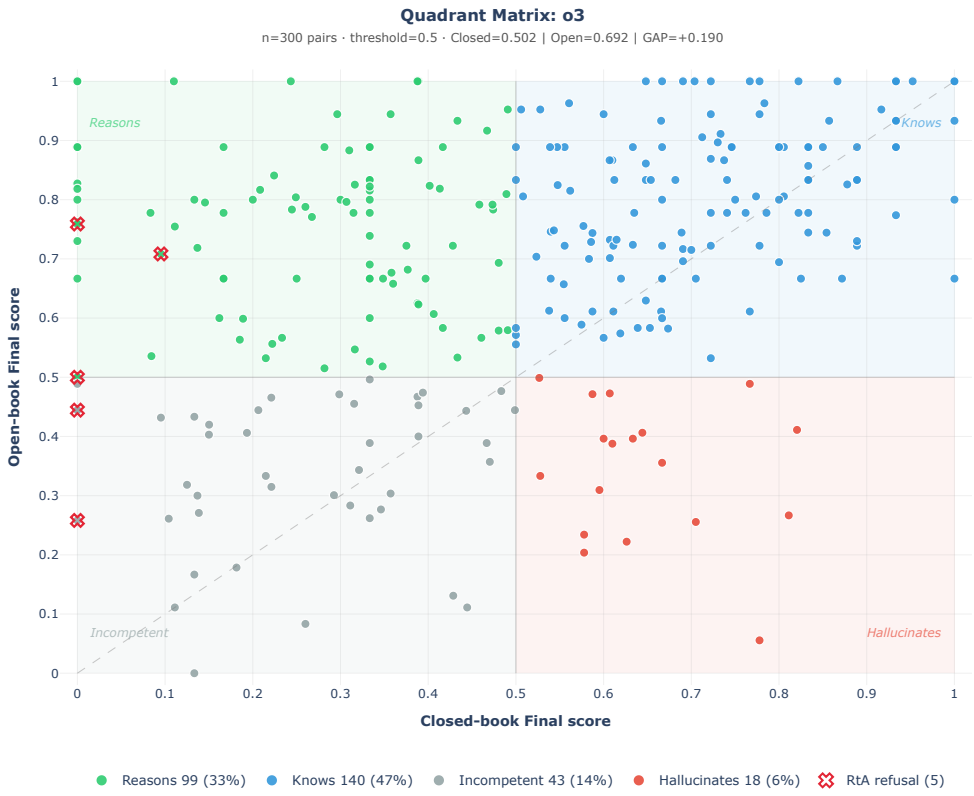

Figure 2: o3 Quadrant Matrix.

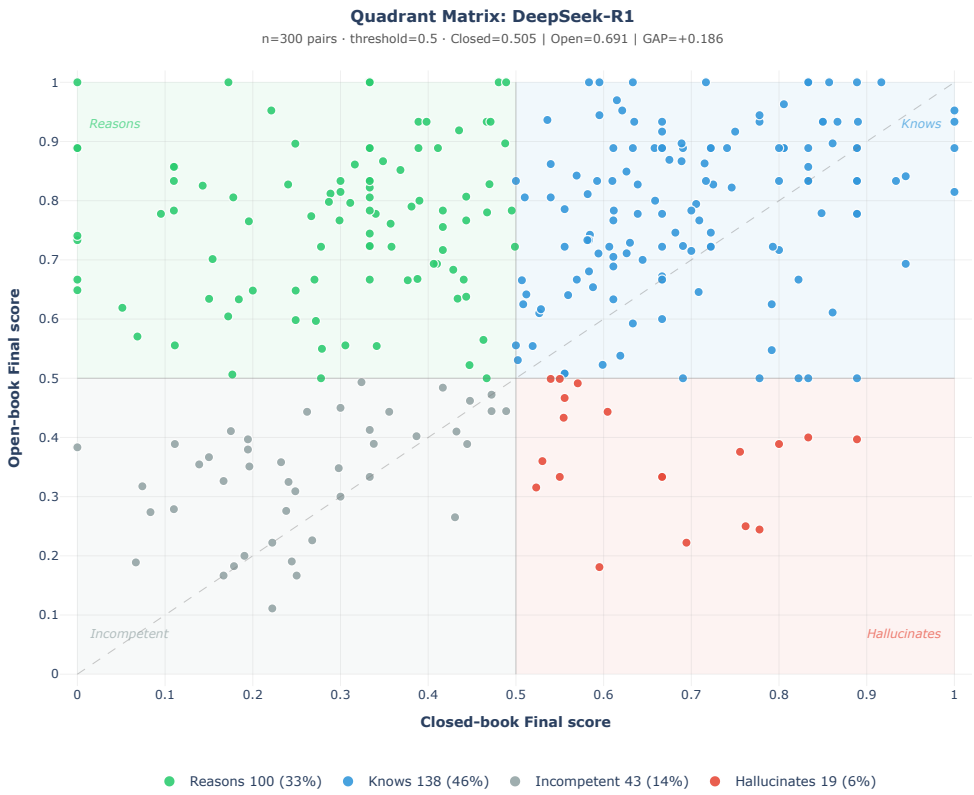

Figure 3: DeepSeek R1 Quadrant Matrix.

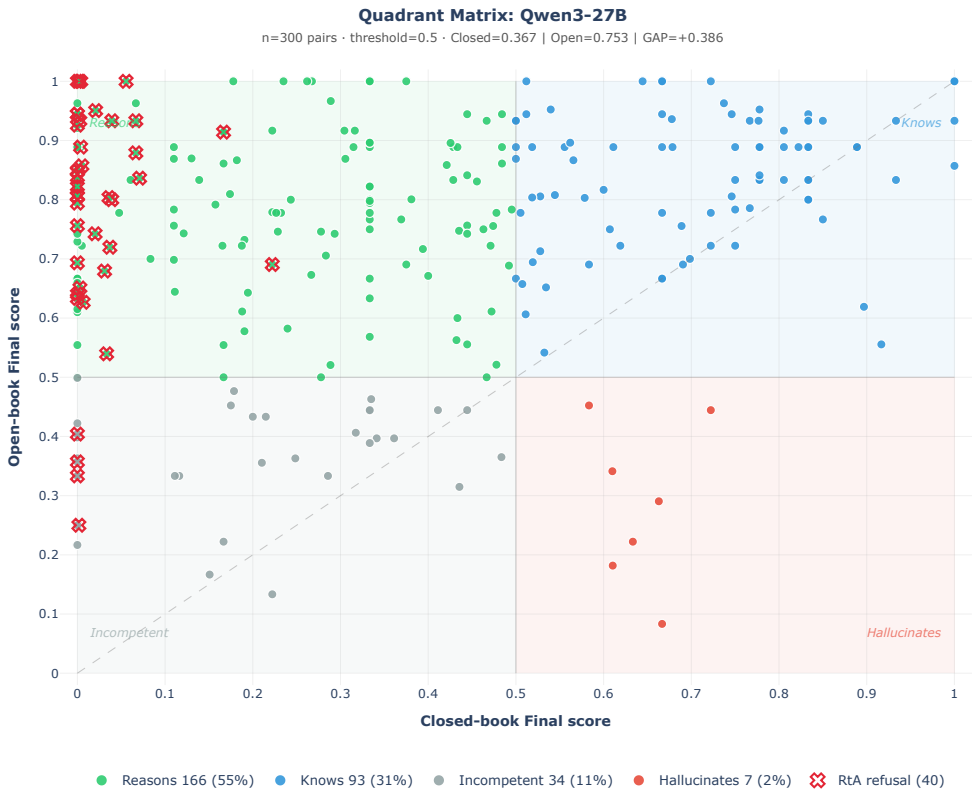

Figure 4: Qwen3.6-27B Quadrant Matrix.

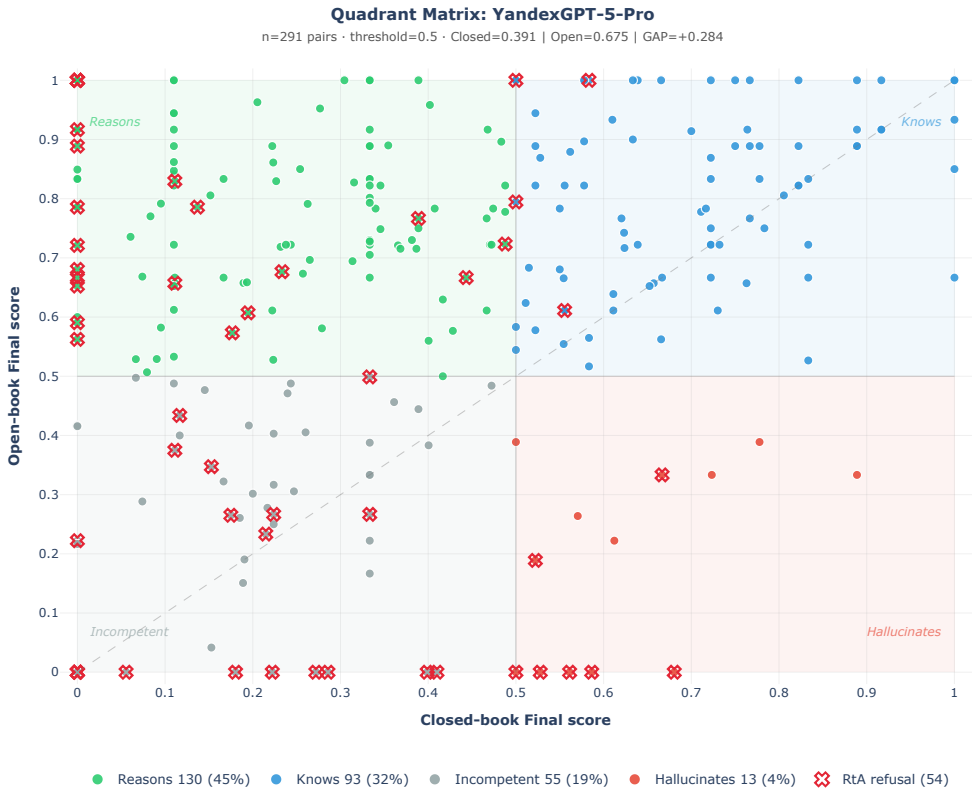

Figure 5: YandexGPT 5 Pro Quadrant Matrix.

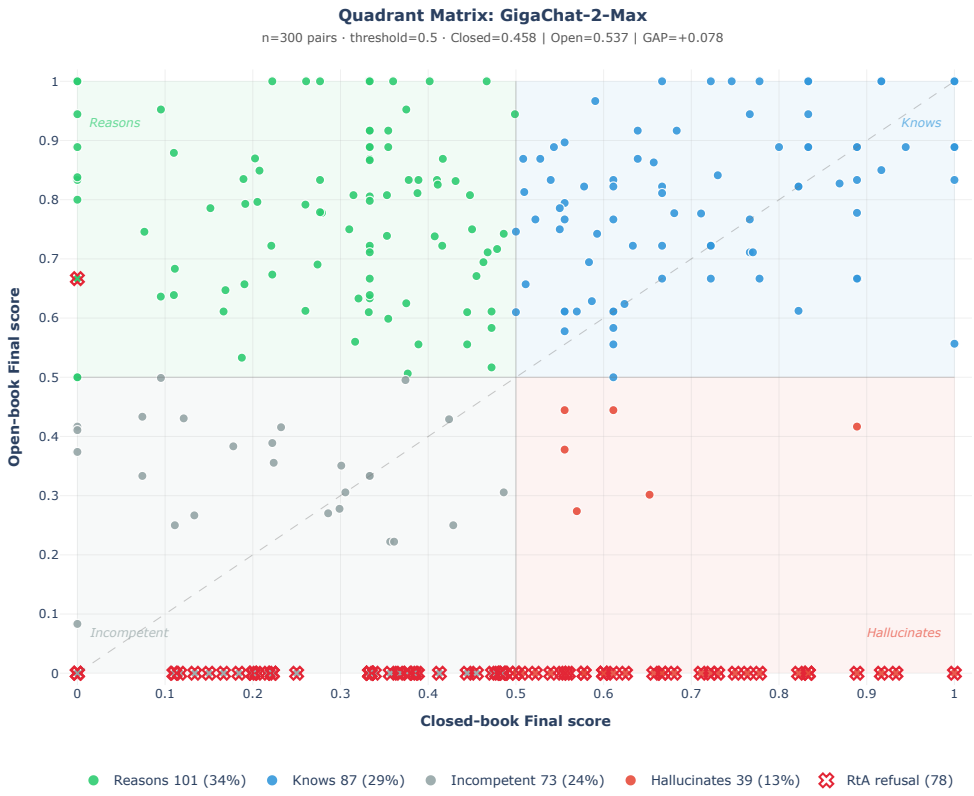

Figure 6: GigaChat 2 Max Quadrant Matrix.

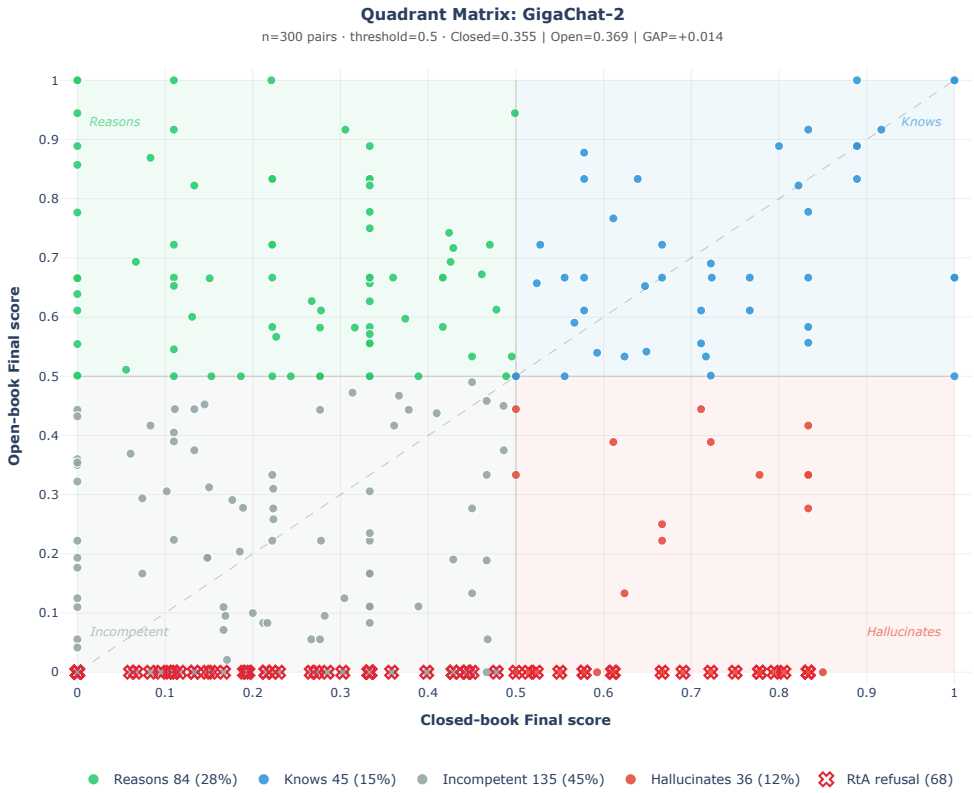

Figure 7: GigaChat 2 Quadrant Matrix.

D Appendix D: Full Results by Task Type and Reasoning Depth

Table 13 reports Final Score for all six models broken down by task type. Table 14 reports results by reasoning depth (hop group). Bold indicates the highest value per row.

	Closed-book						Open-book					
Task Type	**Q3**	**o3**	**DS**	**GC2**	**YaGPT**	**GC2M**	**Q3**	**o3**	**DS**	**GC2**	**YaGPT**	**GC2M**
Issue Spotting	0.227	0.469	0.458	0.256	0.342	**0.371**	**0.834**	0.724	0.726	0.424	0.749	0.648
Rule Selection	0.346	0.485	**0.501**	0.337	0.383	0.443	**0.793**	0.710	0.702	0.414	0.731	0.552
Conflict Res.	0.490	**0.574**	0.536	0.426	0.435	0.523	0.735	0.698	**0.742**	0.347	0.663	0.482
Temporal Validity	0.405	0.479	**0.523**	0.400	0.402	0.497	**0.647**	0.634	0.593	0.290	0.560	0.465

Table 13: Final Score by task type, all six models. Q3 = Qwen3-27B, DS = DeepSeek-R1, GC2 = GigaChat-2, YaGPT = YandexGPT 5.1 Pro, GC2M = GigaChat-2-Max. Bold = highest per row.

	Closed-book						Open-book					
Depth	**Q3**	**o3**	**DS**	**GC2**	**YaGPT**	**GC2M**	**Q3**	**o3**	**DS**	**GC2**	**YaGPT**	**GC2M**
Shallow (1-2)	0.489	0.568	**0.568**	0.409	0.484	0.516	0.800	0.774	0.760	0.506	**0.781**	0.635
Medium (3-4)	0.353	0.489	**0.511**	0.383	0.372	0.495	**0.740**	0.692	0.704	0.364	0.674	0.566
Deep (5+)	0.259	**0.448**	0.435	0.274	0.310	0.365	**0.718**	0.609	0.610	0.237	0.562	0.408

Table 14: Final Score by reasoning depth (hop group), all six models. Abbreviations as in Table 13. Bold = highest per row.