Evaluating Legal Reasoning Traces with Legal Issue Trees

Jinu Lee, Kyoung-Woon On, Simeng Han, Arman Cohan, Julia Hockenmaier

UIUC, LBOX, Stanford, Yale

1/12/2026

Contents

- Introduction
- LEGIT dataset
- Experimental results
  - Reliability of LLM-as-a-judge
  - Performance of LLMs
  - Comparing RAG and RL performance in LEGIT
- Conclusion

Introduction

Two main motivations:

Reasoning trace evaluation is more than final answer evaluation.

- For expert-level reasoning (ex. **law**), users rely on LLM’s reasoning traces.
- In legal reasoning, both **correctness AND completeness** of the arguments are critical
  - Completeness was not often captured in math/STEM reasoning tasks

Legal reasoning field lacks **high coverage & high quality datasets**

- No legal reasoning dataset have evaluated the intermediate reasoning traces
- Existing datasets mostly cover crimes, lacking civil/administrative cases
  - Criminal cases are more like regression; civil cases are more boolean prediction

Introduction

Goals:

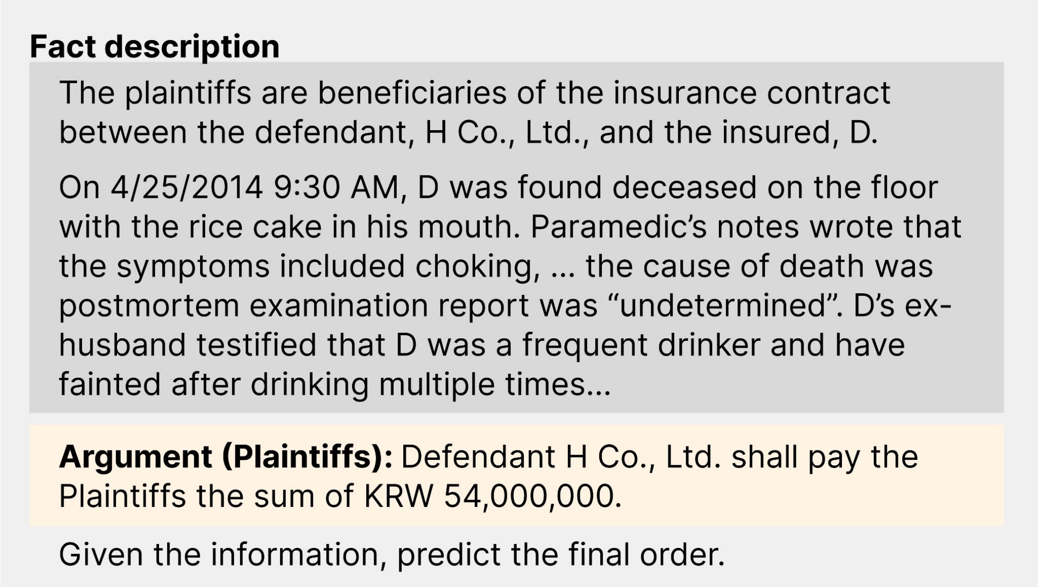
1. Develop a legal reasoning dataset **with reliable/scalable CoT evaluation.**

2. **Understand and improve LLMs' legal reasoning abilities** using the dataset**:**
   - How do SOTA LLMs perform in legal reasoning?
   - How do different methods (RAG, RL, ...) improve LLM's ability?

Introduction

Task: Legal Judgment Prediction

- Given the facts and claims, predict the final judgment of the court.



Fact description

The plaintiffs are beneficiaries of the insurance contract between the defendant, H Co., Ltd., and the insured, D.

On 4/25/2014 9:30 AM, D was found deceased on the floor with the rice cake in his mouth. Paramedic's notes wrote that the symptoms included choking, ... the cause of death was postmortem examination report was "undetermined". D's ex-husband testified that D was a frequent drinker and have fainted after drinking multiple times...

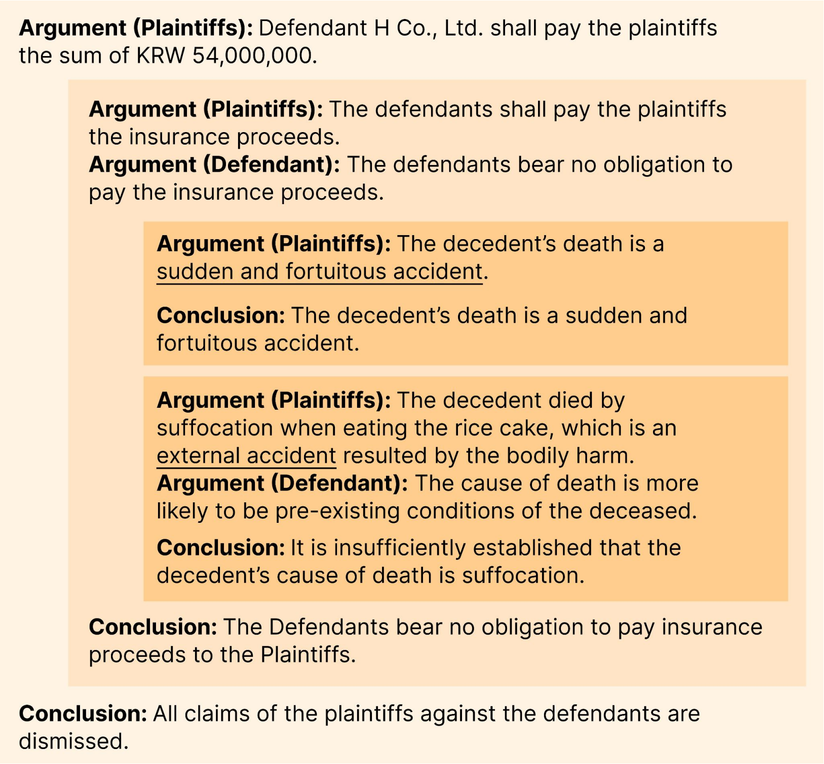
Argument (Plaintiffs): Defendant H Co., Ltd. shall pay the Plaintiffs the sum of KRW 54,000,000.

Given the information, predict the final order.

Introduction

Legal judgment prediction = reasoning through **hierarchical legal issue trees**

Example case:



> **Argument (Plaintiffs):** Defendant H Co., Ltd. shall pay the plaintiffs the sum of KRW 54,000,000.
>
> > **Argument (Plaintiffs):** The defendants shall pay the plaintiffs the insurance proceeds.
> > **Argument (Defendant):** The defendants bear no obligation to pay the insurance proceeds.
> >
> > > **Argument (Plaintiffs):** The decedent's death is a sudden and fortuitous accident.
> > >
> > > **Conclusion:** The decedent's death is a sudden and fortuitous accident.
> >
> > > **Argument (Plaintiffs):** The decedent died by suffocation when eating the rice cake, which is an external accident resulted by the bodily harm.
> > > **Argument (Defendant):** The cause of death is more likely to be pre-existing conditions of the deceased.
> > >
> > > **Conclusion:** It is insufficiently established that the decedent's cause of death is suffocation.
> >
> > **Conclusion:** The Defendants bear no obligation to pay insurance proceeds to the Plaintiffs.
>
> **Conclusion:** All claims of the plaintiffs against the defendants are dismissed.

Gist of claim

청구취지 (피고는 돈을 지급하라)

Insurance proceeds

보험금

Sudden and fortuitous

급격하고도 우연한 사고이다.

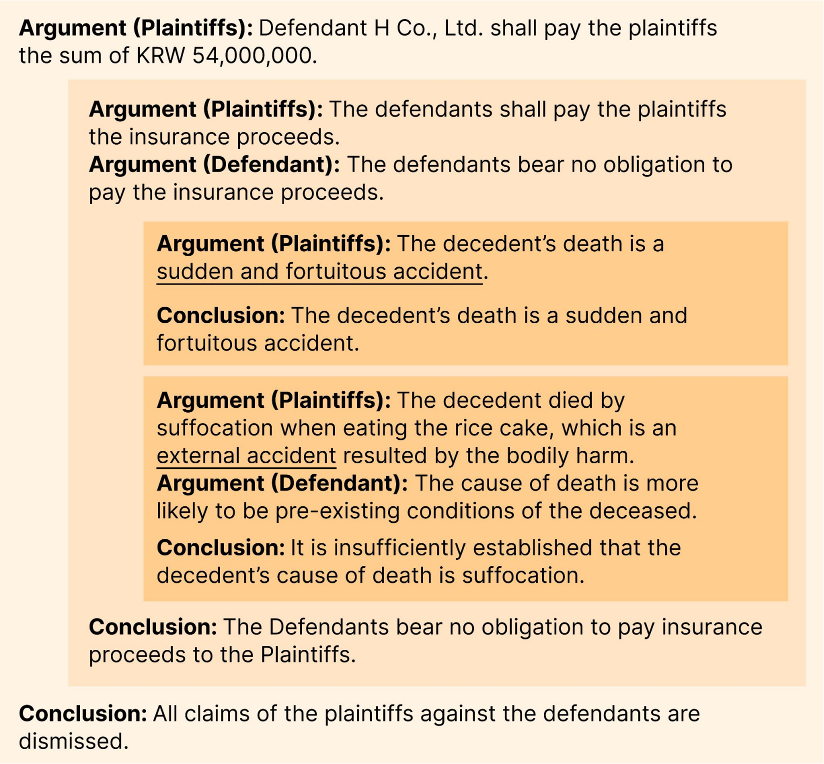
External?

외래의 사고인가?

Introduction

Legal judgment prediction = reasoning through **hierarchical legal issue trees**

Example case:



Argument (Plaintiffs): Defendant H Co., Ltd. shall pay the plaintiffs the sum of KRW 54,000,000.

Argument (Plaintiffs): The defendants shall pay the plaintiffs the insurance proceeds.
Argument (Defendant): The defendants bear no obligation to pay the insurance proceeds.

Argument (Plaintiffs): The decedent's death is a sudden and fortuitous accident.

Conclusion: The decedent's death is a sudden and fortuitous accident.

Argument (Plaintiffs): The decedent died by suffocation when eating the rice cake, which is an external accident resulted by the bodily harm.
Argument (Defendant): The cause of death is more likely to be pre-existing conditions of the deceased.

Conclusion: It is insufficiently established that the decedent's cause of death is suffocation.

Conclusion: The Defendants bear no obligation to pay insurance proceeds to the Plaintiffs.

Conclusion: All claims of the plaintiffs against the defendants are dismissed.

Gist of claim
→ Denied

Is the accident ensured?
→ Denied

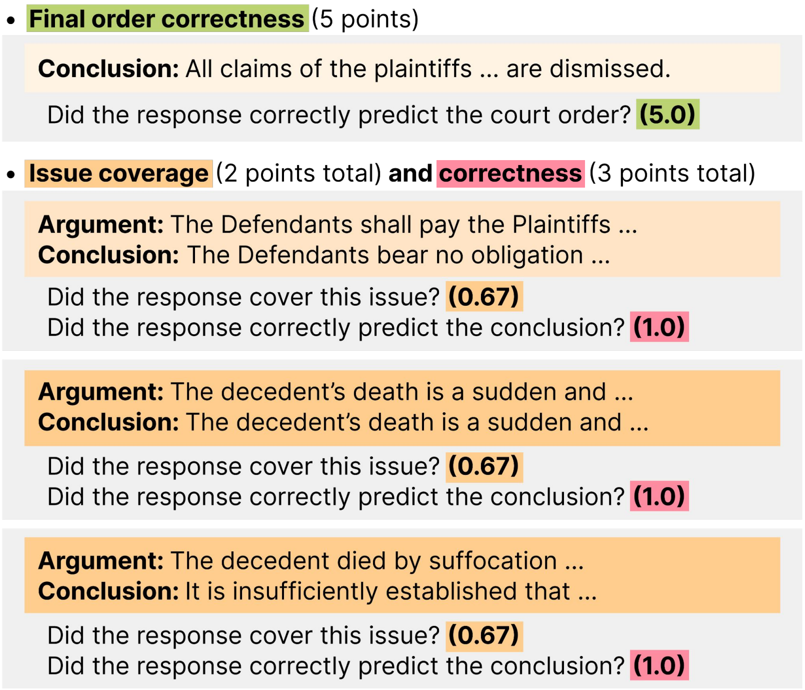
Sudden and fortuitous?
→ Accepted

External?
→ Denied

Introduction

Legal issues are intermediate rewards for evaluating reasoning traces

- Did the LLM **identify** this legal issue? (= Coverage)
- Did the LLM **find the correct conclusion** of this issue? (= Correctness)



- **Final order correctness** (5 points)

Conclusion: All claims of the plaintiffs ... are dismissed.

Did the response correctly predict the court order? **(5.0)**

- **Issue coverage** (2 points total) **and correctness** (3 points total)

Argument: The Defendants shall pay the Plaintiffs ...
Conclusion: The Defendants bear no obligation ...

Did the response cover this issue? **(0.67)**
Did the response correctly predict the conclusion? **(1.0)**

Argument: The decedent's death is a sudden and ...
Conclusion: The decedent's death is a sudden and ...

Did the response cover this issue? **(0.67)**
Did the response correctly predict the conclusion? **(1.0)**

Argument: The decedent died by suffocation ...
Conclusion: It is insufficiently established that ...

Did the response cover this issue? **(0.67)**
Did the response correctly predict the conclusion? **(1.0)**

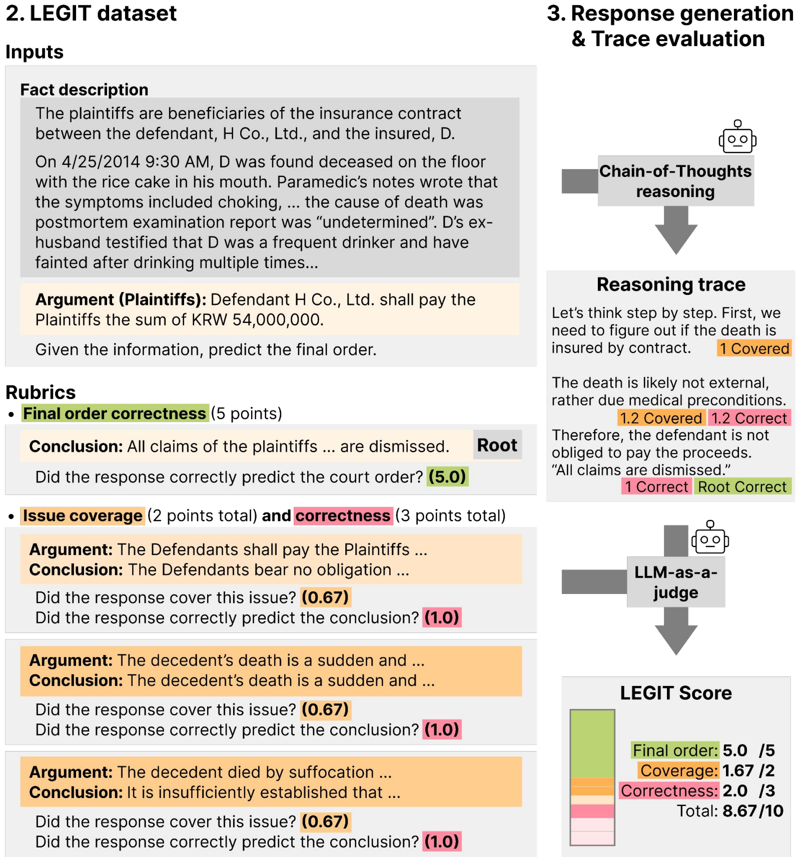
= Final answer reward

= Intermediate reward

Introduction

Legal issue rubrics evaluate reasoning traces

- **Generator LLM** solves legal judgment prediction with Chain-of-thoughts
- **LLM-as-a-judge** evaluates CoT with legal issue rubrics



2. LEGIT dataset

Inputs

Fact description

The plaintiffs are beneficiaries of the insurance contract between the defendant, H Co., Ltd., and the insured, D.

On 4/25/2014 9:30 AM, D was found deceased on the floor with the rice cake in his mouth. Paramedic's notes wrote that the symptoms included choking, ... the cause of death was postmortem examination report was "undetermined". D's ex-husband testified that D was a frequent drinker and have fainted after drinking multiple times...

Argument (Plaintiffs): Defendant H Co., Ltd. shall pay the Plaintiffs the sum of KRW 54,000,000.

Given the information, predict the final order.

Rubrics

- **Final order correctness** (5 points)

Conclusion: All claims of the plaintiffs ... are dismissed. **Root**

Did the response correctly predict the court order? **(5.0)**

- **Issue coverage** (2 points total) **and correctness** (3 points total)

Argument: The Defendants shall pay the Plaintiffs ...
Conclusion: The Defendants bear no obligation ...

Did the response cover this issue? **(0.67)**
Did the response correctly predict the conclusion? **(1.0)**

Argument: The decedent's death is a sudden and ...
Conclusion: The decedent's death is a sudden and ...

Did the response cover this issue? **(0.67)**
Did the response correctly predict the conclusion? **(1.0)**

Argument: The decedent died by suffocation ...
Conclusion: It is insufficiently established that ...

Did the response cover this issue? **(0.67)**
Did the response correctly predict the conclusion? **(1.0)**

3. Response generation & Trace evaluation

Chain-of-Thoughts reasoning

Reasoning trace

Let's think step by step. First, we need to figure out if the death is insured by contract. 1 Covered

The death is likely not external, rather due medical preconditions. 1.2 Covered 1.2 Correct

Therefore, the defendant is not obliged to pay the proceeds. "All claims are dismissed." 1 Correct Root Correct

LLM-as-a-judge

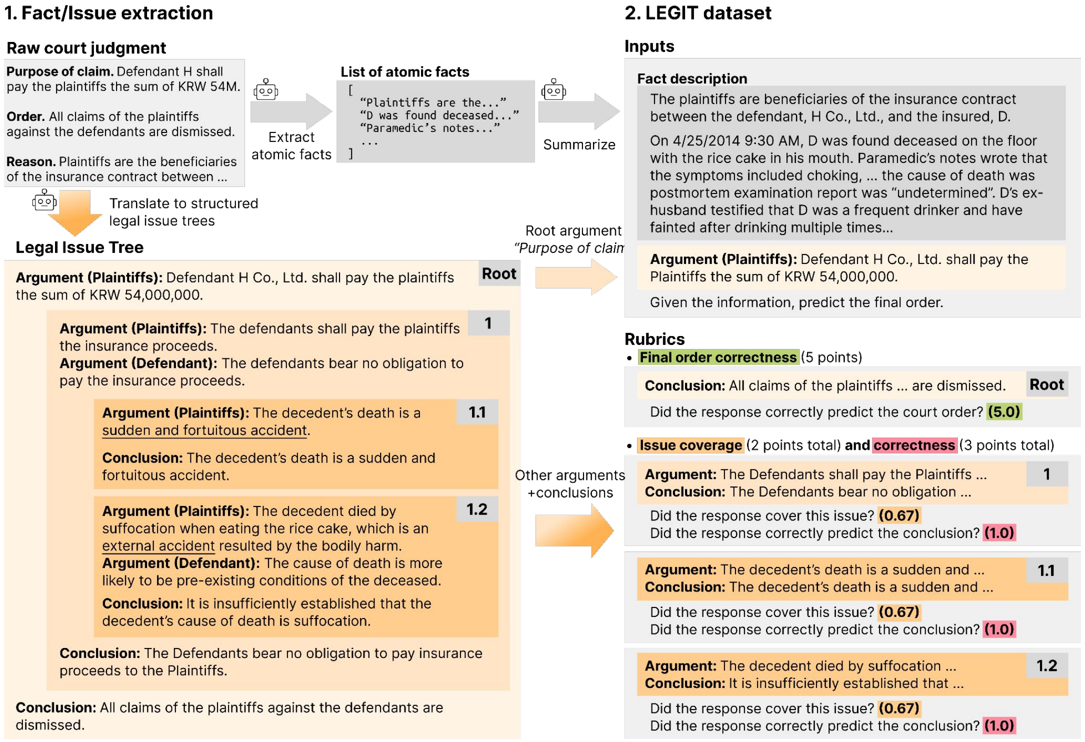
LEGIT Score

Final order: 5.0 /5
Coverage: 1.67 /2
Correctness: 2.0 /3
Total: 8.67/10

LEGIT dataset

LEGIT dataset:

- Start from raw court judgment
- LEGIT inputs
  - Extract atomic facts
  - Compose into single text
- LEGIT rubrics
  - Translate judgments to issue tree using LLMs
  - Convert issues to rubrics



1. Fact/Issue extraction

Raw court judgment

Purpose of claim. Defendant H shall pay the plaintiffs the sum of KRW 54M.

Order. All claims of the plaintiffs against the defendants are dismissed.

Reason. Plaintiffs are the beneficiaries of the insurance contract between ...

Extract atomic facts

List of atomic facts

```
[
  “Plaintiffs are the...”
  “D was found deceased...”
  “Paramedic’s notes...”
  ...
]
```

Summarize

Translate to structured legal issue trees

Legal Issue Tree

Root — **Argument (Plaintiffs):** Defendant H Co., Ltd. shall pay the plaintiffs the sum of KRW 54,000,000.

1 — **Argument (Plaintiffs):** The defendants shall pay the plaintiffs the insurance proceeds. **Argument (Defendant):** The defendants bear no obligation to pay the insurance proceeds.

1.1 — **Argument (Plaintiffs):** The decedent’s death is a sudden and fortuitous accident.

Conclusion: The decedent’s death is a sudden and fortuitous accident.

1.2 — **Argument (Plaintiffs):** The decedent died by suffocation when eating the rice cake, which is an external accident resulted by the bodily harm. **Argument (Defendant):** The cause of death is more likely to be pre-existing conditions of the deceased.

Conclusion: It is insufficiently established that the decedent’s cause of death is suffocation.

Conclusion: The Defendants bear no obligation to pay insurance proceeds to the Plaintiffs.

Conclusion: All claims of the plaintiffs against the defendants are dismissed.

Root argument “Purpose of clai

Other arguments +conclusions

2. LEGIT dataset

Inputs

Fact description

The plaintiffs are beneficiaries of the insurance contract between the defendant, H Co., Ltd., and the insured, D.

On 4/25/2014 9:30 AM, D was found deceased on the floor with the rice cake in his mouth. Paramedic’s notes wrote that the symptoms included choking, ... the cause of death was postmortem examination report was “undetermined”. D’s ex-husband testified that D was a frequent drinker and have fainted after drinking multiple times...

Argument (Plaintiffs): Defendant H Co., Ltd. shall pay the Plaintiffs the sum of KRW 54,000,000.

Given the information, predict the final order.

Rubrics

- **Final order correctness** (5 points)

Root — **Conclusion:** All claims of the plaintiffs ... are dismissed.

Did the response correctly predict the court order? **(5.0)**

- **Issue coverage** (2 points total) **and correctness** (3 points total)

1 — **Argument:** The Defendants shall pay the Plaintiffs ... **Conclusion:** The Defendants bear no obligation ...

Did the response cover this issue? **(0.67)**
Did the response correctly predict the conclusion? **(1.0)**

1.1 — **Argument:** The decedent’s death is a sudden and ... **Conclusion:** The decedent’s death is a sudden and ...

Did the response cover this issue? **(0.67)**
Did the response correctly predict the conclusion? **(1.0)**

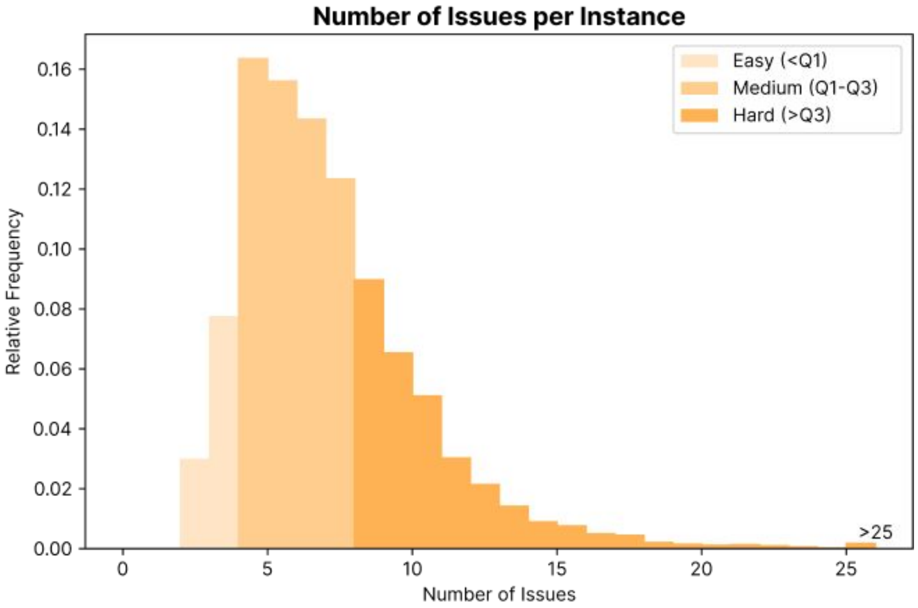
1.2 — **Argument:** The decedent died by suffocation ... **Conclusion:** It is insufficiently established that ...

Did the response cover this issue? **(0.67)**
Did the response correctly predict the conclusion? **(1.0)**

LEGIT dataset

Dataset statics

- **~24.5k** diverse Korean district court cases (civil & administration)
- Average issues per instance: **7.3**



Experiments & Results

LEGIT: Legal reasoning dataset that evaluates issue coverage and correctness of CoT.

RQ 1. Can LLM-as-a-judge reliably evaluate reasoning traces with LEGIT's rubrics?

RQ 2. How do LLMs perform in legal reasoning?

RQ 3. How do RAG and RL improve legal reasoning performance?

Experiments & Results: RQ 1

RQ 1. Can LLM-as-a-judge effectively evaluate legal reasoning traces with LEGIT's rubrics?

Results outline:

- LEGIT rubrics allow significant expert-LLM agreement.
- LEGIT rubrics allow more consistent evaluation than underspecified rubrics.

Experiments & Results: RQ 1

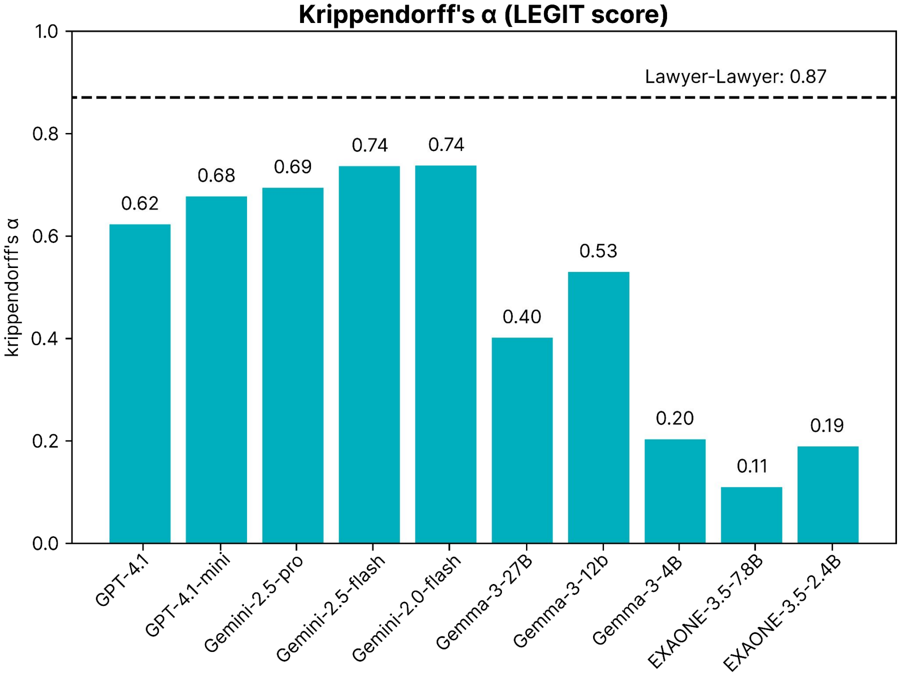
Experiment: Measure human-LLM/LLM-LLM agreement with LEGIT rubrics.

- Using the test set (300 examples),
- **Generate responses for 12 LLMs** (that can fluently generate long Korean text)
    - OpenAI GPT-4.1, GPT-4.1-mini, o3
    - Google Claude 2.5-Pro, 2.5-Flash, 2.0-Flash
    - Gemma-3-{4B, 12B, 27B}
    - EXAONE-3.5-{32B, 8B}, EXAONE-3.0-8B
- **Evaluate all solutions (300*12) with 11 different LLMs**
    - OpenAI GPT-4.1, GPT-4.1-mini
    - Google Claude 2.5-Pro, 2.5-Flash, 2.0-Flash
    - Gemma-3-{4B, 12B, 27B}
    - EXAONE-3.5-{8B, 2.4B}, EXAONE-3.0-8B
- Collect 2 Korean lawyers' annotations on LEGIT rubrics
    - 44 responses, 300 issues

Experiments & Results: RQ 1

Strong LLMs achieve significant agreement with human experts

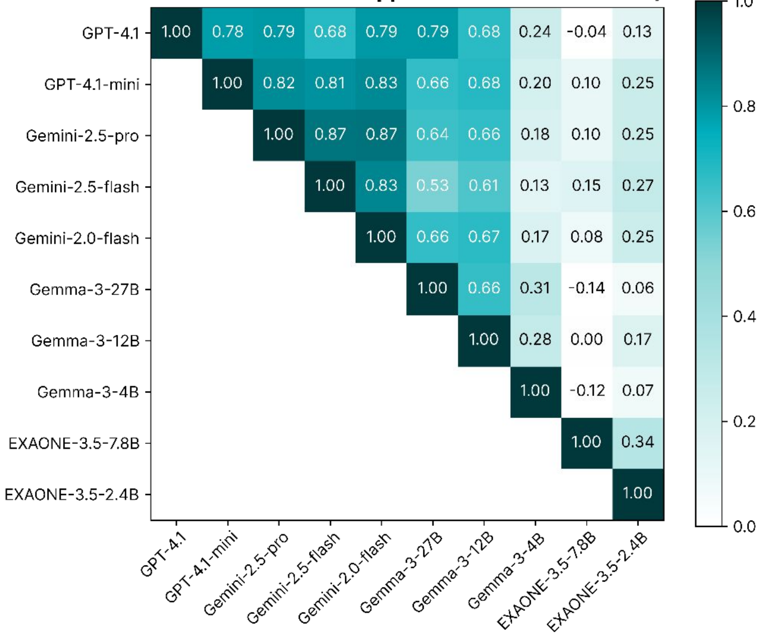
- Lawyer and strong (closed-source) LLMs show significant agreement (α>0.67).
- Gemma-12B and 27B show moderately high agreement,
  while other OW models are close to random



Experiments & Results: RQ 1

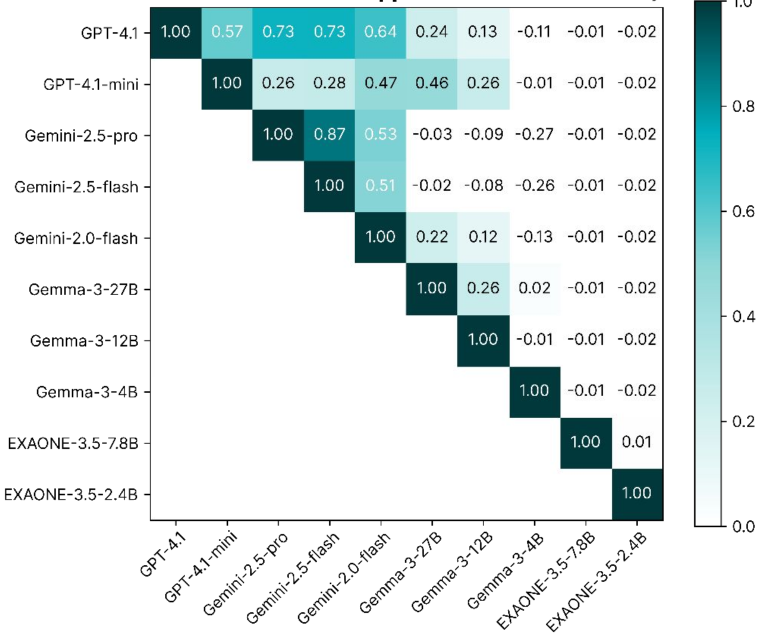
LLM evaluations are more consistent with LEGIT rubrics than other rubrics

- **Control group: Likert scale** (0-10), with (1) score descriptions and (2) raw judgments
- LLM-LLM agreement is substantially higher with LEGIT rubrics than Likert scale.
- **LEGIT rubrics allow more consistent evaluation!**



LLM-LLM Pairwise Krippendorff's α (LEGIT score)

	GPT-4.1	GPT-4.1-mini	Gemini-2.5-pro	Gemini-2.5-flash	Gemini-2.0-flash	Gemma-3-27B	Gemma-3-12B	Gemma-3-4B	EXAONE-3.5-7.8B	EXAONE-3.5-2.4B
GPT-4.1	1.00	0.78	0.79	0.68	0.79	0.79	0.68	0.24	-0.04	0.13
GPT-4.1-mini		1.00	0.82	0.81	0.83	0.66	0.68	0.20	0.10	0.25
Gemini-2.5-pro			1.00	0.87	0.87	0.64	0.66	0.18	0.10	0.25
Gemini-2.5-flash				1.00	0.83	0.53	0.61	0.13	0.15	0.27
Gemini-2.0-flash					1.00	0.66	0.67	0.17	0.08	0.25
Gemma-3-27B						1.00	0.66	0.31	-0.14	0.06
Gemma-3-12B							1.00	0.28	0.00	0.17
Gemma-3-4B								1.00	-0.12	0.07
EXAONE-3.5-7.8B									1.00	0.34
EXAONE-3.5-2.4B										1.00



LLM-LLM Pairwise Krippendorff's α (Likert scale)

	GPT-4.1	GPT-4.1-mini	Gemini-2.5-pro	Gemini-2.5-flash	Gemini-2.0-flash	Gemma-3-27B	Gemma-3-12B	Gemma-3-4B	EXAONE-3.5-7.8B	EXAONE-3.5-2.4B
GPT-4.1	1.00	0.57	0.73	0.73	0.64	0.24	0.13	-0.11	-0.01	-0.02
GPT-4.1-mini		1.00	0.26	0.28	0.47	0.46	0.26	-0.01	-0.01	-0.02
Gemini-2.5-pro			1.00	0.87	0.53	-0.03	-0.09	-0.27	-0.01	-0.02
Gemini-2.5-flash				1.00	0.51	-0.02	-0.08	-0.26	-0.01	-0.02
Gemini-2.0-flash					1.00	0.22	0.12	-0.13	-0.01	-0.02
Gemma-3-27B						1.00	0.26	0.02	-0.01	-0.02
Gemma-3-12B							1.00	-0.01	-0.01	-0.02
Gemma-3-4B								1.00	-0.01	-0.02
EXAONE-3.5-7.8B									1.00	0.01
EXAONE-3.5-2.4B										1.00

Experiments & Results: RQ 2

RQ 2. How do LLMs perform in legal reasoning?

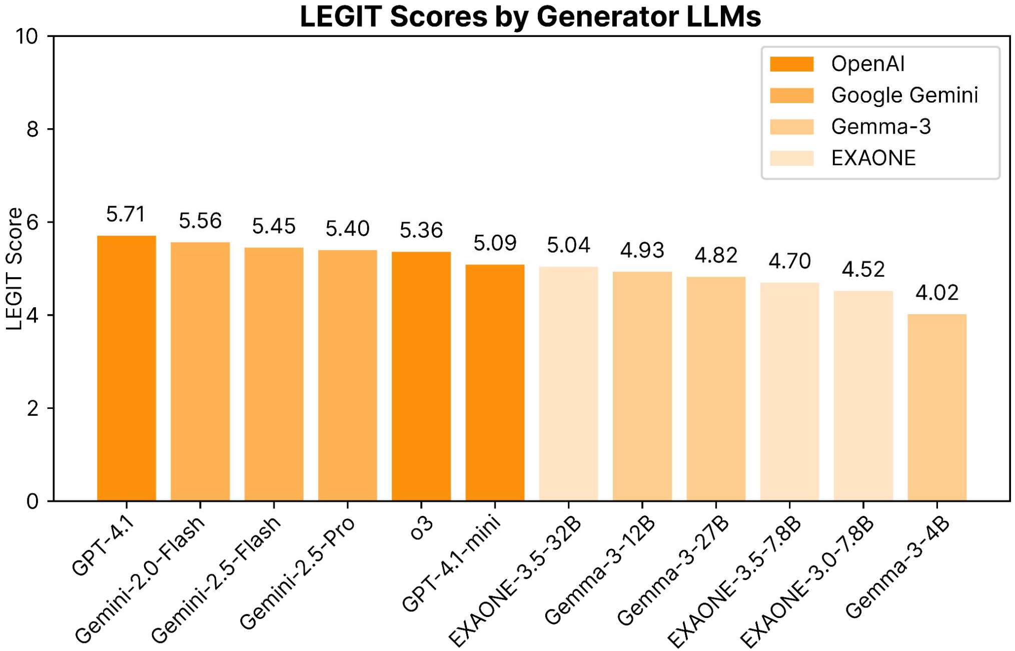
Results outline:
(Now that we trust LLM-as-a-judge scores,)

- LLMs often fail to:
  (1) identify relevant issues **(low coverage)**
  (2) predict their outcome **(low correctness)**
- These errors affect the answer accuracy and reasoning trace quality.

Experiments & Results: RQ 2

Even the strongest LLMs are not perfect

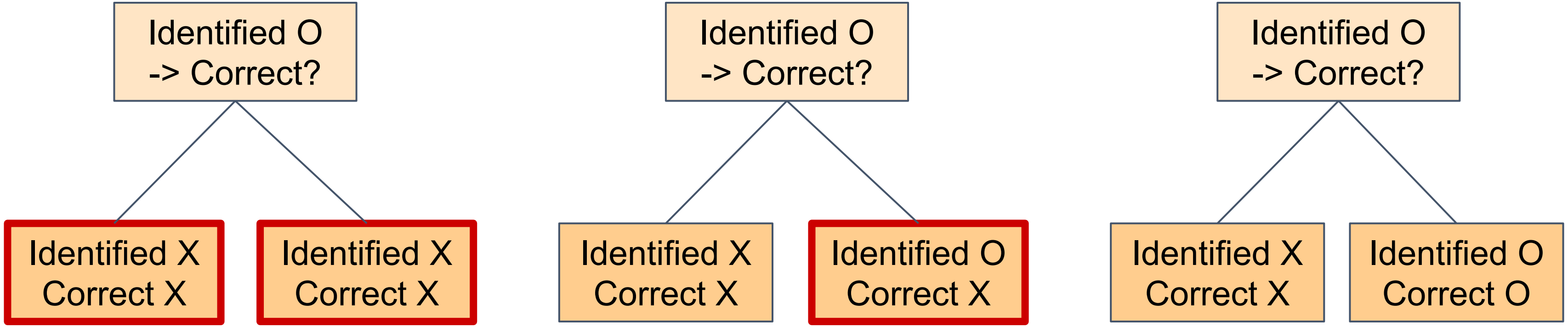
- Best LLM: GPT-4.1 (5.71/10)
- Interesting observation: o3 is worse than non-reasoning models
  - primarily due to low final answer correctness



Experiments & Results: RQ 2

Children issue coverage and correctness both affect the parent issue correctness.

- Analyzing depth-1 subtrees:



Case 1.
No children identified

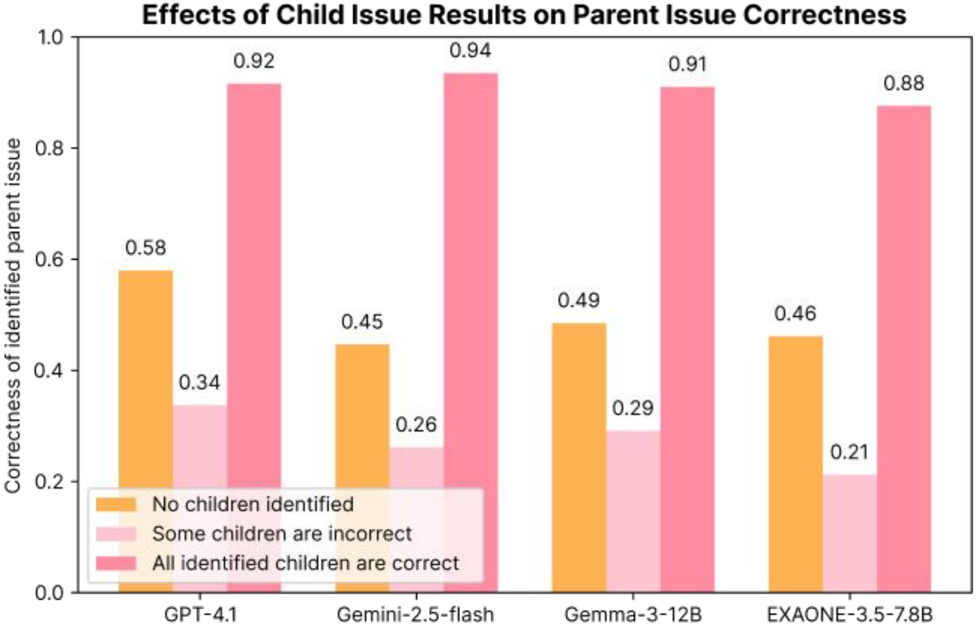
Case 2.
Some children are identified but incorrect

Case 3.
All identified children are correct

Experiments & Results: RQ 2

Low child issue coverage and correctness propagates to parent issue correctness.

- Parent node correctness:
  - Failing to identify child issues: GPT-4.1 **92% -> 58%**
  - Errors in child issues: GPT-4.1 **92% -> 34%**



Experiments & Results: RQ 3

RQ 3. How do RAG and RL improve legal reasoning performance?

Results outline:

- **RAG** improves **all reasoning abilities**
  - Performace improves despite low retrieval performance
- **RL** significantly improves **final order/issue correctness** but reduces issue coverage
  - RL focuses on reducing uncertainty by deliberately ignoring vague issues

Experiments & Preliminary results: RQ 3

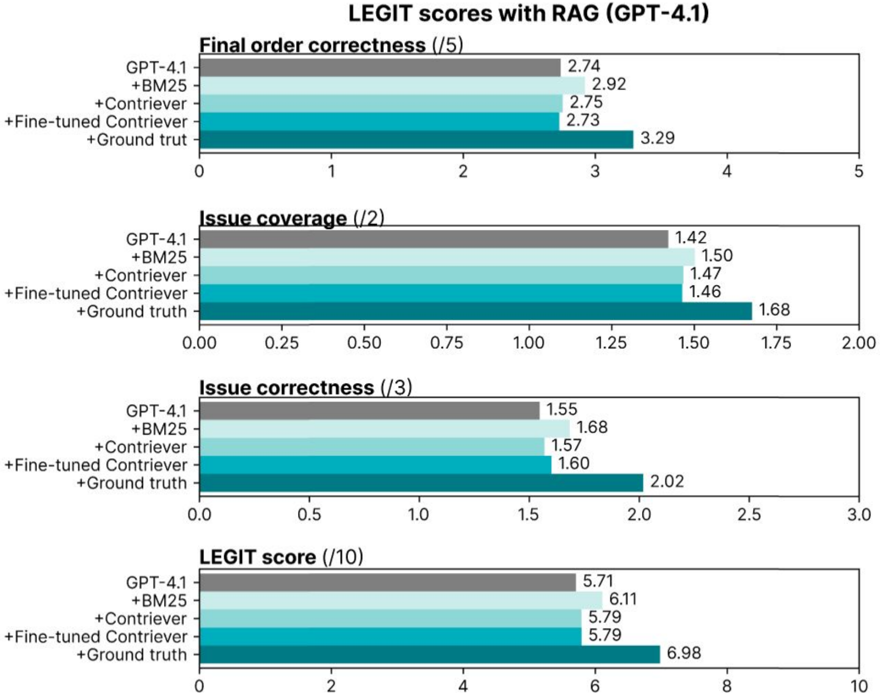
RAG with citation retrieval

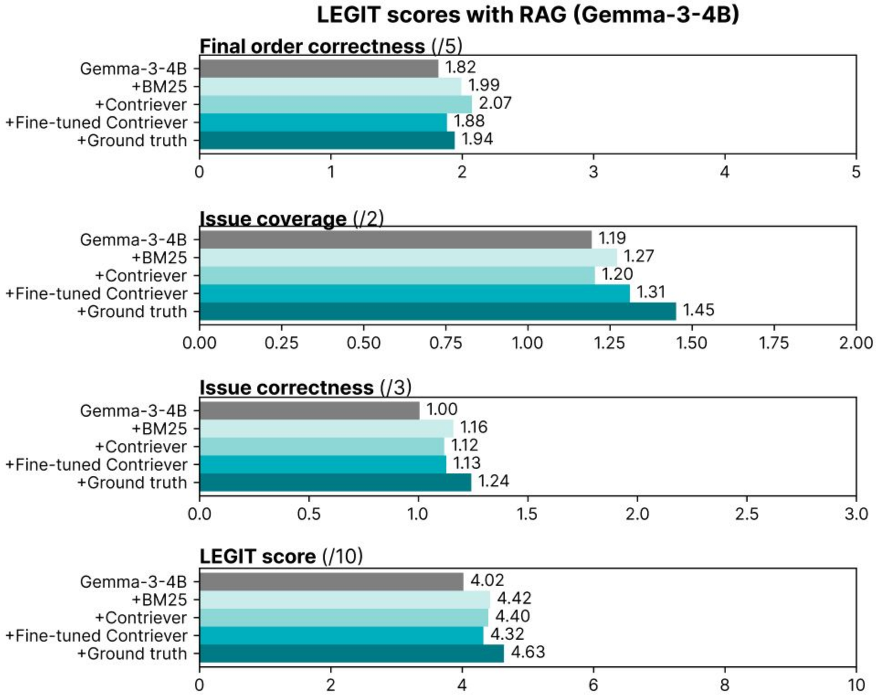
- **Retrieval query:** LEGIT query (facts, gist of claim)
- **Retrieval base:** relevant law (statute + cases) cited in LEGIT training/test set
- **Retrieval methods:**
  - No RAG
  - BM25 (keyword matching)
  - Contriever (vector-based retrieval)
  - Ground-truth citations (extracted from original judgment)
- **Response generation:** Prepend top 10 search results to reasoner LLM's context

Experiments & Preliminary results: RQ 3

RAG improves all reasoning abilities (final answer correctness, issue coverage, issue correctness)

- Ground-truth citations bring significant gain to all three sub-scores
- While other retrievers' performance are low, LLMs can still benefit from RAG





Experiments & Preliminary results: RQ 3

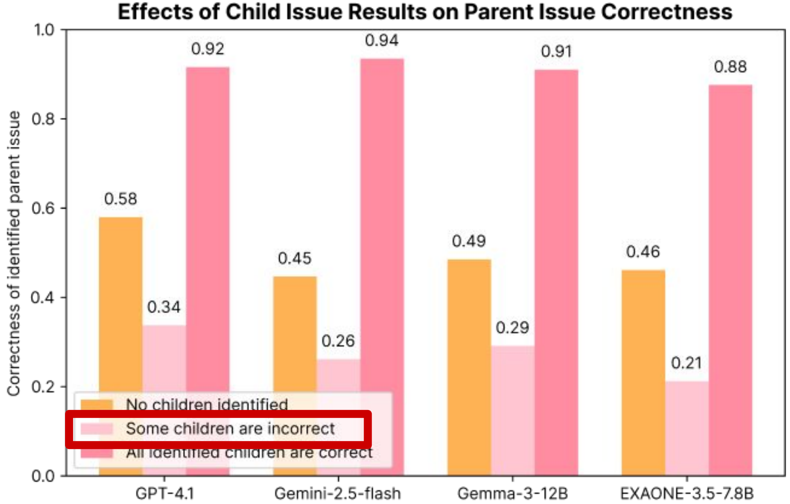
RL with rubrics-as-rewards

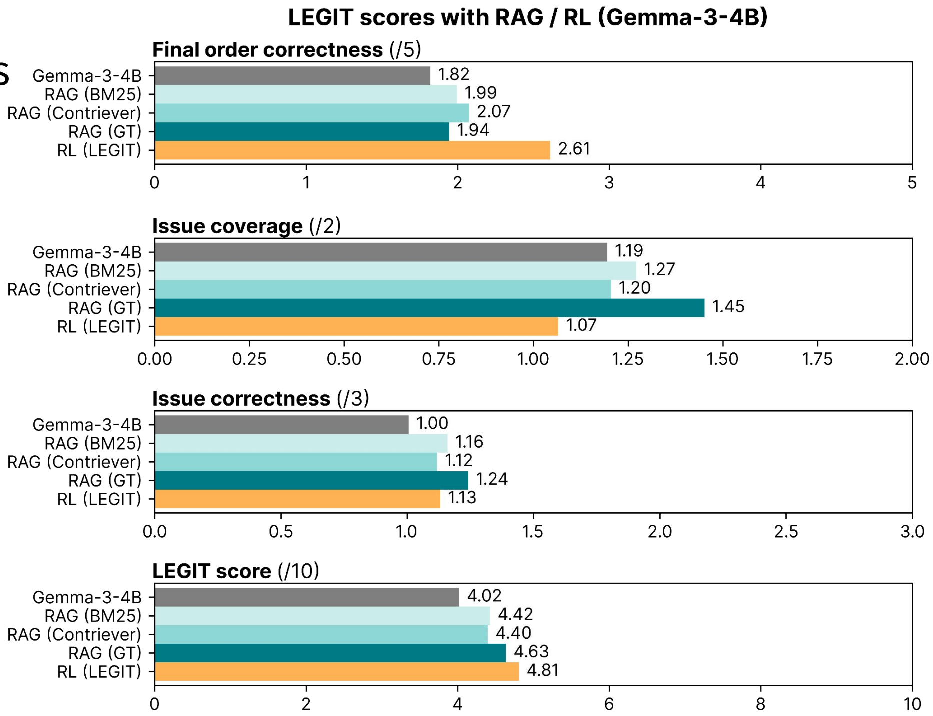
- Use LEGIT score as RL rewards
  - 0-10 points, evaluated with LEGIT rubrics and LLM-as-a-judge
  - -5 points if including format error (n-gram repetition, code-switching)
- Settings:
  - Policy: **Gemma-3-4b-it**
  - LLM-as-a-judge during training: **Gemma-3-27b-it**
  - LLM-as-a-judge during evaluation: **Gemini-2.0-Flash**

Experiments & Preliminary results: RQ 3

Results

- RL dramatically improves correctness, but reduces issue coverage
- Given that “covered but incorrect” issues severely affect parent issues, the policy learns to not mention vague issues that might misguide its reasoning

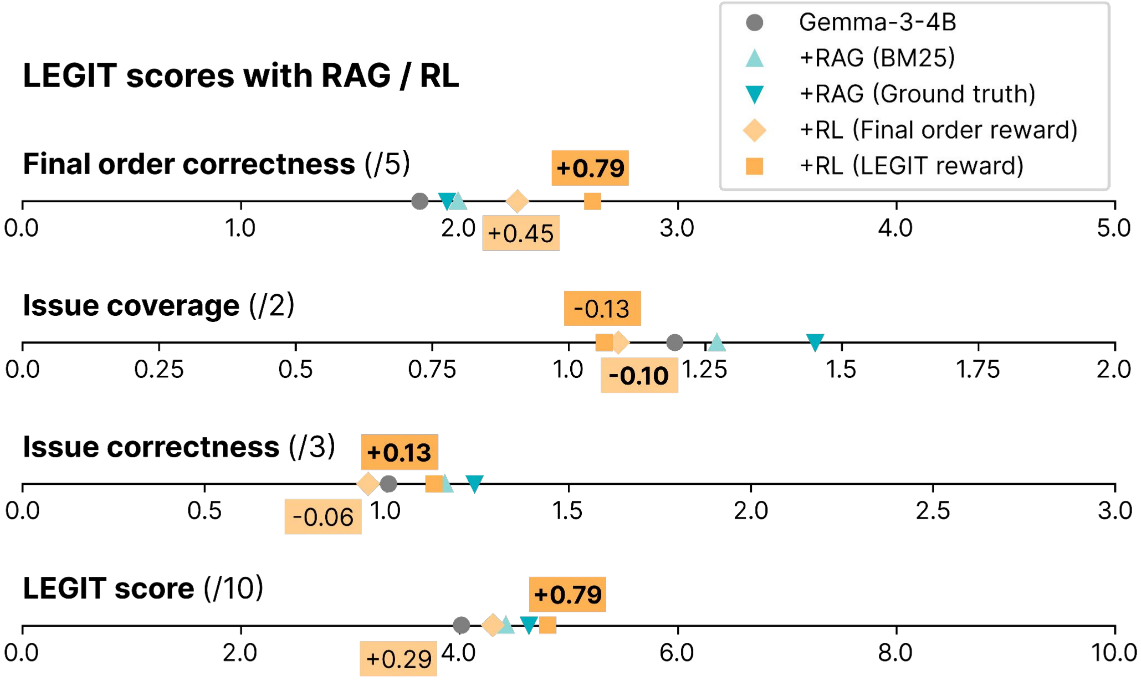




Experiments & Preliminary results: RQ 3

Bonus: comparing with final answer-only rewards

- Train with **reward = final answer correctness** (ignore issue coverage/correctness)
- The resulting model is suboptimal compared to LEGIT score reward.



Conclusion

RQ 1. Can LLM-as-a-judge reliably evaluate reasoning traces with LEGIT's rubrics?
A. LEGIT rubrics allow high expert-LLM agreement and more consistent evaluation.

RQ 2. How do LLMs perform in legal reasoning?
A. Even SOTA LLMs expose shortcomings in issue coverage/correctness.
Error (fail to identify/reason) in child issues propagate to parent issues.

RQ 3. How do RAG and RL improve legal reasoning performance?
A. RAG improves all capabilities, even with limited retrieval performance.
RL improves correctness and reduces coverage to minimize error propagation.

Combination of domain-specific structures (issue trees) and LLM-as-a-judge allows better evaluation and improvement of expert-level reasoning tasks!