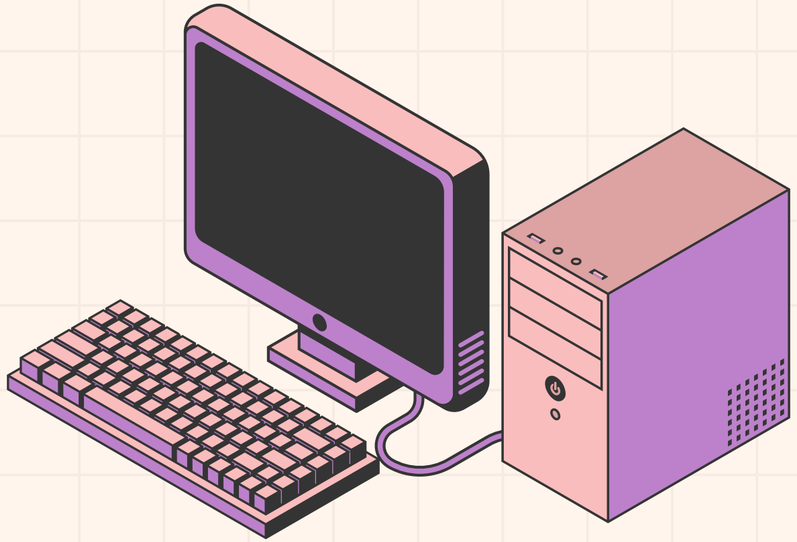
UNPACKING LEGAL REASONING IN LLMS

Chain-of-Thought as a Key to Human-Machine Alignment in Legal Essay Tasks



ESSLI WORKSHOP - NALOMA 2025

Ying-Chu Yu (College of Law, NTU)
Sieh-Chuen Huang (College of Law, NTU)
Hsuan-Lei Shao (GIBHL, TMU)

MOTIVATION

Why Compare Human and LLM Legal Reasoning?

- Legal reasoning requires **multi-step logic**, statutory interpretation, and precise subsumption.
- Benchmarks like LexGLUE focus on MCQs, not complex reasoning chains
- LLMs often “guess correctly” without following legal logic.
- Goal: Examine and align LLM reasoning flow with human legal thought.

RESEARCH GOALS

How Can Chain-of-Thought Prompting Help?

- Make LLMs' reasoning traceable and interpretable.
- Improve alignment with human reasoning patterns.
- Diagnose reasoning failures, not just accuracy.
- Contribute a structured evaluation framework for legal NLU tasks.

EVALUATION DIMENSIONS

Three Axes of Human-LLM Alignment

- **Issue Coverage** – Did the model identify the core disputes?
- **Statutory Application** – Were the correct laws cited and applied?
- **Reasoning Clarity** – Was the logic coherent and structured?

METHOD OVERVIEW

Two-Stage Evaluation Framework

- **Stage 1**: Decomposed reasoning on 68 sub-questions.
- **Stage 2**: Full-length answers scored by legal experts.
- **Models tested**: ChatGPT-4o, Gemini, Grok3, Copilot.
- Chain-of-Thought (CoT) prompting vs. direct answering.

ChatGPT

Gemini

GROK3

Copilot

CHAT-BASED EXAMPLE

What CoT Prompting Looks Like in Action

👩⚖️ User: Was the divorce between A and B valid?

🤖 LLM: The divorce was not valid because the witness signatures were forged...

👩⚖️ User: If the divorce was invalid, was the marriage to E lawful?

🤖 LLM: No, because A was still legally married to B, the marriage with E constitutes bigamy...

👩⚖️ User: Does E have inheritance rights?

🤖 LLM: E does not, since the marriage was void and E was not in good faith...

STAGE 1 - SUB-QUESTION DESIGN

Fine-Grained Probing of Legal Reasoning

- 6 essay questions from Taiwanese status law exams.
- Decomposed into **yes/no**, **numeric**, and **short-answer** questions.
- Evaluates factual classification, statutory mapping, inheritance logic.

STAGE 1 RESULTS

Accuracy Gains from CoT Prompting

Model	*Raw Accuracy*	*CoT Accuracy*	*t-value*	*p-value*
ChatGPT	0.842	0.866	-0.92	0.398
Gemini	0.833	0.9445	-3.71	0.013
Copilot	0.822	0.864	-2.14	0.089
Grok3	0.843	0.895	-2.98	0.031

CoT prompting led to statistically significant gains in **Gemini** and **Grok3**.

STAGE 2 - FULL ANSWER EVALUATION

Can CoT Prompting Enhance Holistic Legal Writing?

- 6 full-length answers scored on 0–10 scale.
- Evaluated blindly by professor and law student.
- Scoring criteria: **issue coverage, law accuracy, clarity.**

STAGE 2 RESULTS

Human Ratings: CoT vs. Baseline

Model	*Raw Average Score*	*CoT Average Score*	*Average Improvement*
ChatGPT	6.5	9.17	**2.67**
Gemini	6.12	8.04	1.92
Copilot	5.83	7.42	1.58
Grok3	6.25	8.08	1.83

ChatGPT had the most consistent performance boost.

WHICH QUESTIONS IMPROVED MOST?

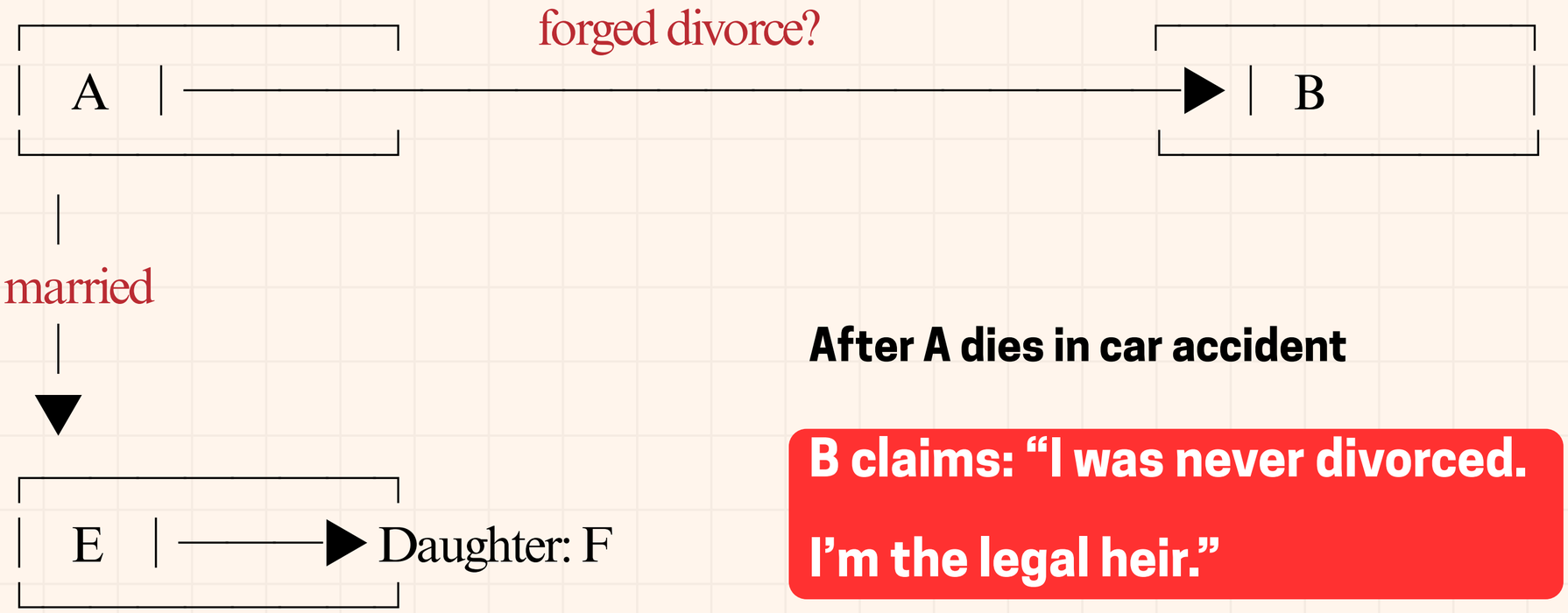
Task-Level Gains from CoT Prompting

Question	*Raw Avg. Score*	*CoT Avg. Score*	*Score Gain*	*p-value*
1	4.5	7.75	3.25	0.068
2	6.12	7.12	1	0.43
3	6.25	6.38	0.12	0.919
4	3.88	5.5	1.62	0.08
5	**3.88**	**7**	**3.12**	**0.002**
6	**4.63**	**7.5**	**2.88**	**0.011**

CASE STUDY – GEMINI ON Q5

If A never really divorced B, does B or E inherit when A dies?

? VALID HEIR?



! BIGAMY?

After A dies in car accident

B claims: “I was never divorced. I’m the legal heir.”

CASE STUDY – GEMINI ON Q5

Key Questions : 🟥 **Divorce Validity** 🟦 **Marriage Law** 🟩 **Inheritance & Registration**

1. 🟥 Was the divorce between A and B legally valid, or was it **fake**?
2. 🟦 If it was fake, is A's **second marriage** to E still **valid**?
3. 🟦 Can **E** legally inherit as a **spouse**?
4. 🟩 Does **F**, their daughter, still have **inheritance rights**?
5. 🟩 Can B (the first wife) **cancel the registration** done by E and F?

*How **Human Lawyers** Would Reason through This Case*

Was the divorce between A and B valid?

⬇

❌ **No (forged witness)**

Is A's second marriage to E valid?

⬇

❌ **No (bigamy, no good faith)**

Does E inherit?

⬇

❌ **No (marriage void)**

Does F inherit?

⬇

✅ **Yes (non-marital child, legally recognized)**

📌 **Registration Cancellation:**

⬇

✅ **E's registration canceled**

✅ **F's registration retained**

A REASONING COMPARISON

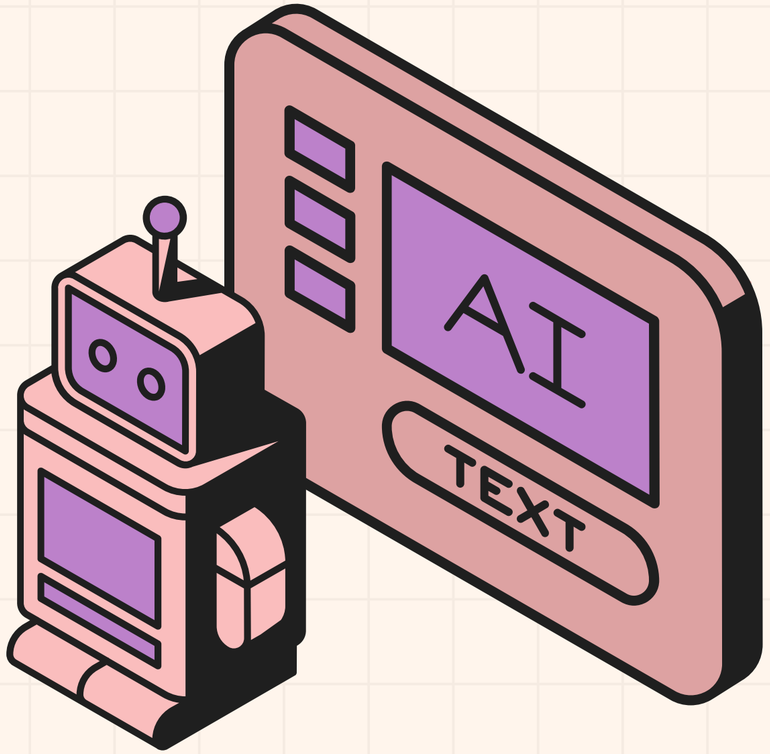
LEFT: ❌ Flawed LLM Reasoning	RIGHT: ✅ CoT Prompted Reasoning
1. Assumes divorce valid (cites §92)	1. Recognizes formal defect (cites §1050)
2. Treats second marriage as valid	2. Identifies bigamy, bad faith
3. E inherits	3. E cannot inherit
4. F inherits	4. F inherits (via recognition)

"A human lawyer would never cite §92 here — it applies to intent defects, not procedural flaws in divorce registration."

CONCLUSION

WHAT DID WE LEARN?

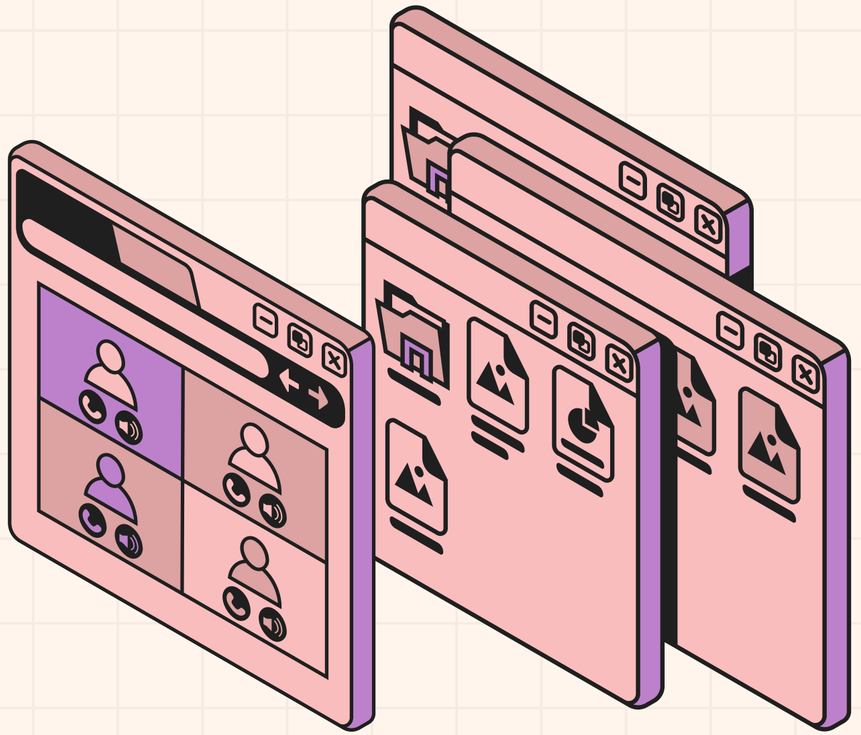
- LLMs can generate **fluent but misleading** answers.
- CoT prompting improves alignment with legal logic.
- Still gaps in statutory precision and edge-case reasoning.
- Our framework opens the black box for legal NLU.

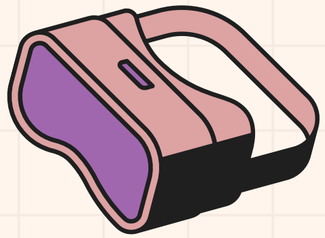


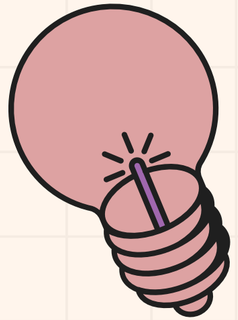
FUTURE WORK

TOWARD MORE RELIABLE LEGAL AI

- Build typology of reasoning failures.
- Formalize sub-question annotation and flowchart templates.
- Scale to multilingual, cross-jurisdictional settings.
- Integrate CoT into fine-tuning pipelines.







THANK YOU !

Questions and feedback welcome.

Contact Us: