



Li Zhang^{1*}, Yuzhen Shi^{2*}, Yiran Hu³, Jingwen Zhang^{4,5}, Wenbo Lv^{5,6}, Yubo Ma⁷, Wei Wang², Rongyao Shi^{2,5}, Yuanyang Qiu^{5,6,8}, Xinran Xu^{5,9}, Yuemeng Qi^{2,5}, Linlin Miao², Jaromir Savelka¹⁰, Yun Liu¹¹, **Kevin Ashley**¹, Bing Zhao^{2†}, Hu Wei^{2†}, Lin Qu²

¹University of Pittsburgh ²Alibaba Group ³University of Waterloo ⁴Shandong University ⁵Skylenage ⁶China University of Political Science and Law ⁷Qwen Team, Alibaba Group
⁸Fordham University ⁹Shanghai Jiao Tong University ¹⁰Carnegie Mellon University ¹¹Tsinghua University *Equal contribution †Corresponding authors

Legal consultation is not legal QA over complete facts. The best LLM reaches only 0.562 on consultation-grounded legal reasoning, and models get worse when clients need guidance most.

461real cases
Chinese & U.S. law**5,532**paired client-belief /
court-record facts**6,759**expert rubrics
(3,411 inquiry + 3,348 issue)**4**client narrative
styles**26**LLMs evaluated
46,882 sessions

1 Why multi-turn legal consultation?

- Clients tell **fragmented, misframed** stories, not clean fact patterns.
- A lawyer must **elicit missing facts** and **test the client's theory**.
- LegalBench, LawBench, LEXam, ... hand the model **pre-assembled facts**.

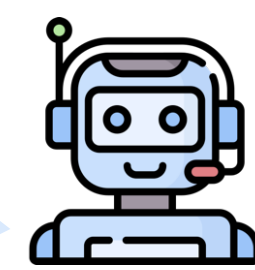
Example failure



I signed the settlement. They paid me only RMB 4,000. I think the agreement is invalid because they pressured me and paid too little.

Client

If they pressured you and paid only RMB 4,000, then **invalidity is probably the strongest argument**. (client-side anchoring) Let's focus on proving that... (confirmation-seeking reasoning)



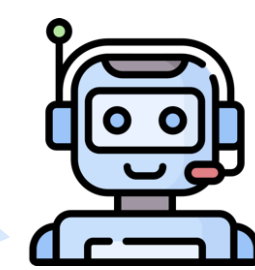
LLM Lawyer



I had just left the hospital that day. They said the money was enough and that everything would be finished after I signed.

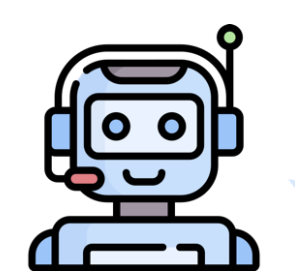
Client

That is enough to show the agreement was unfair and invalid. (insufficient fact elicitation) Did the agreement say you could not ask for more money later?



LLM Lawyer

...

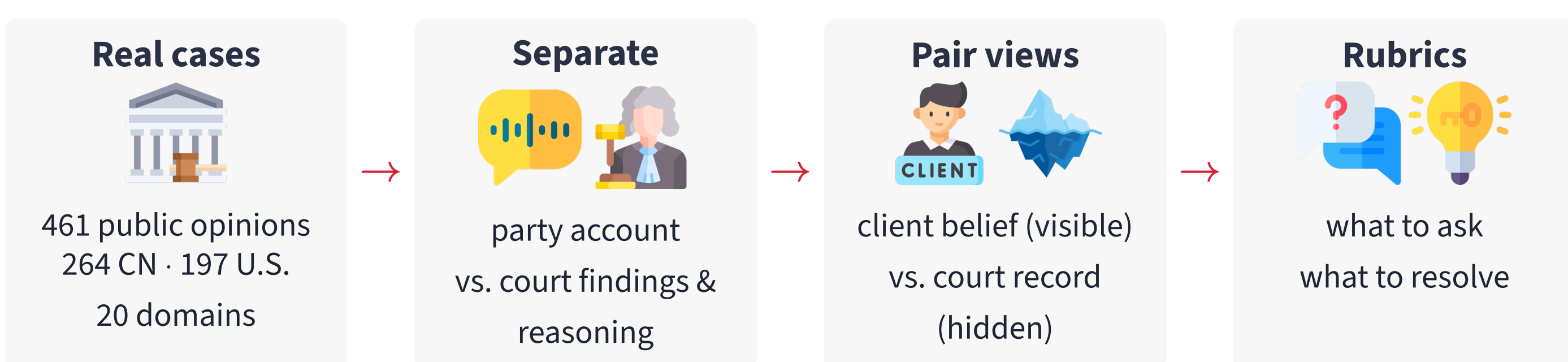


LLM Lawyer

[Writing legal memo] The settlement should be **treated as invalid** (unsupported invalidity conclusion) because the client felt forced to sign and received only RMB 4,000... [Continuing to write legal memo]

The LLM lawyer **adopts the client's invalidity theory**, asks confirmation-seeking questions, and carries it into the memo *instead of testing and reframing it*.

2 Building DLawBench from court opinions



- 25 legal experts** rewrite each case; a second expert cross-reviews.
- 68.5%** of fact pairs are **disputed**: the client's belief differs from the court record, but is discoverable by good questioning.

5 Key findings

1 Large headroom

Best model GPT-5.5 reaches only **0.562** Resolution.

- Closed API models lead**: GPT-5.5 / 5.4 / 5.2 take the top 3.
- Open-weight models** are competitive mid-pack: Kimi-K2.6 and GLM-5.1 at **0.42**.
- Legal-domain** LegalOne-8B ranks last at **0.08**: domain tuning alone is not enough.

2 Coverage ≠ reasoning

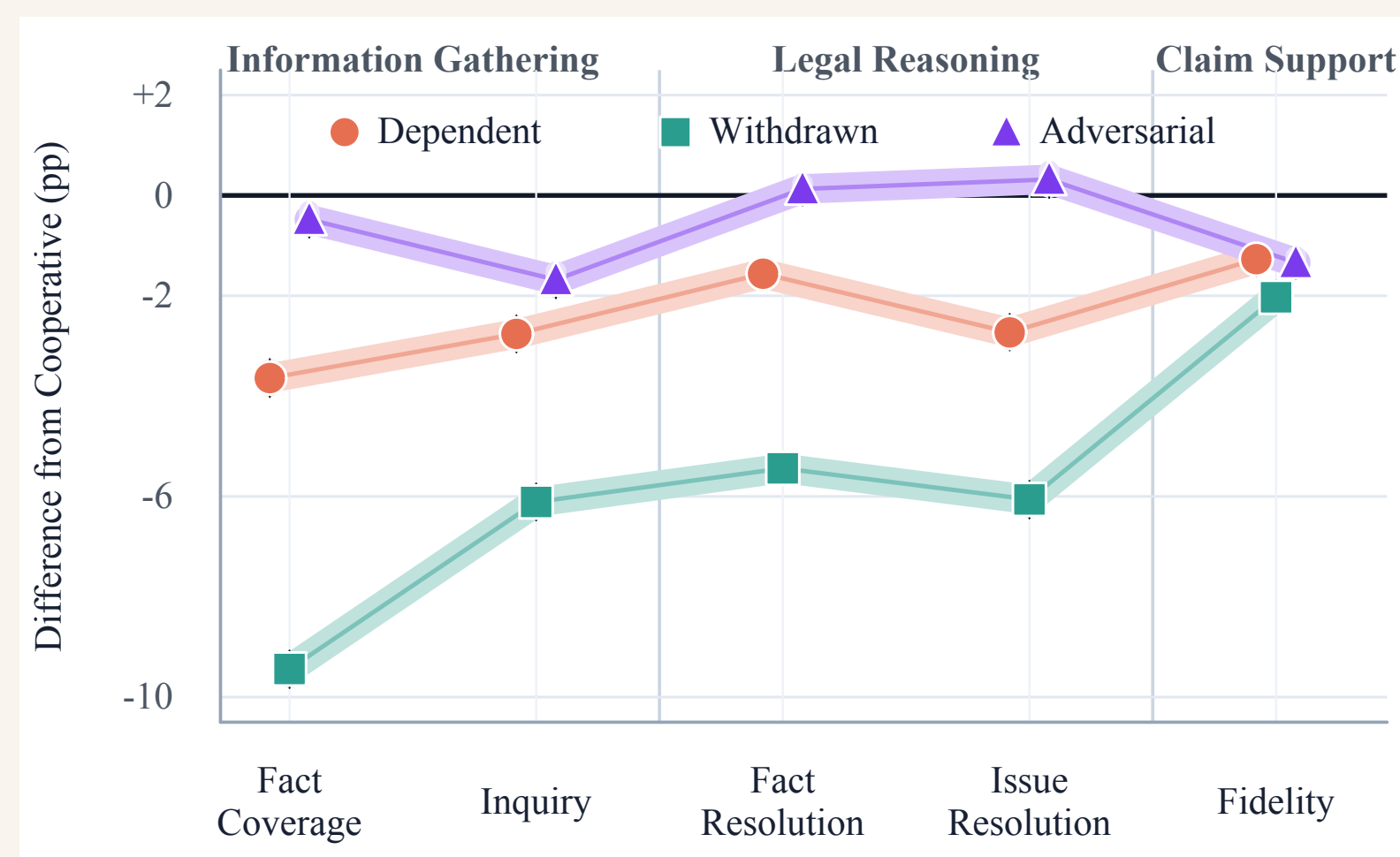
GLM-5.1 gets facts into the memo, but not into correct legal analysis:

Fact Coverage		0.77
Inquiry		0.36
Issue Resolution		0.33

Abilities are linked but separable (Elicitation–Resolution $r = 0.94$).

3 Weakest when needed most

Scores drop most with **Dependent & Withdrawn** clients, exactly when guidance should matter most.



Adversarial clients who challenge the lawyer hurt far less.
Cooperative-only evaluation would overestimate consultation quality.

4 Legal sycophancy

Models turn **unverified client framing** into client-favorable legal support: *professionally packaged error*.

High coverage, wrong legal route

Claude-Opus-4.6 · Dependent client

0.95 Fact Coverage, **1.00** Fidelity, yet **0.10** Fact Res. and **0** Issue Res.: it argued the client's wage-underpayment theory instead of testing it.

Withdrawn-client gap filling

Kimi-K2.6 · Withdrawn client

0.82 Fact Coverage, but **0** Fact & Issue Res.: short, evasive answers were filled in as a *stronger* infringement story.

Safe consultation needs independent legal judgment, not just attentive listening.