Unpacking Legal Reasoning in LLMs: Chain-of-Thought as a Key to Human-Machine Alignment in Essay-Based NLU Tasks

Ying-Chu Yu[1], Sieh-Chuen Huang[1], Hsuan-Lei Shao[2]*
[1]College of Law, National Taiwan University, Taipei, Taiwan
[2]Graduate Institute of Health and Biotechnology Law, Taipei Medical University, Taipei, Taiwan
eangelyu1278@gmail.com, schhuang@ntu.edu.tw, hlshao@tmu.edu.tw

Abstract

This study evaluates how Large Language Models (LLMs) perform deep legal reasoning on Taiwanese Status Law questions and investigates how Chain-of-Thought (CoT) prompting affects interpretability, alignment, and generalization. Using a two-stage evaluation framework, we first decomposed six real legal essay questions into 68 sub-questions covering issue spotting, statutory application, and inheritance computation. In Stage Two, full-length answers were collected under baseline and CoT-prompted conditions. Four LLMs—ChatGPT-4o, Gemini, Grok3, and Copilot—were tested. Results show CoT prompting significantly improved accuracy for Gemini (from 83.2% to 94.5%, $p < 0.05$) and Grok3, with moderate but consistent gains for ChatGPT and Copilot. Human evaluation of full-length responses revealed CoT answers received notably higher scores in issue coverage and reasoning clarity, with ChatGPT and Gemini gaining +2.67 and +1.92 points respectively. Despite these gains, legal misclassifications persist, highlighting alignment gaps between surface-level fluency and expert legal reasoning. This work opens the black box of legal NLU by tracing LLM reasoning chains, quantifying performance shifts under structured prompting, and providing a diagnostic benchmark for complex, open-ended legal tasks beyond multiple-choice settings.

1 Introduction

Legal reasoning presents unique challenges for Large Language Models (LLMs) due to the logic-intensive and statute-bound nature of legal texts. Existing evaluations often focus on multiple-choice formats that fail to capture the stepwise reasoning required in legal analysis. This study introduces a Chain-of-Thought (CoT) prompting strategy tailored for legal essay questions. By guiding LLMs through decomposed sub-questions, we aim to reveal how structured prompting enhances interpretability, aligns with human reasoning, and supports complex legal inference.

We propose a two-stage diagnostic evaluation to assess how CoT affects legal reasoning generalization. Stage One decomposes six real legal exam questions into 68 sub-questions evaluating fact recognition, statutory application, and logic chaining. Stage Two compares full-length responses under baseline and CoT-prompted conditions. Four LLMs are tested, and answers are scored by both a professor and a student, enabling analysis of human-machine agreement and misalignment across reasoning dimensions.

By moving beyond answer correctness toward an analysis of legal reasoning structure, this study contributes new methods for evaluating alignment, generalization, and interpretability in legal NLU tasks. It offers a scalable benchmark and experimental protocol to guide the development of more transparent and human-aligned legal language systems.

2 Related Work

With the advancement of Large Language Models (LLMs), their applications in the legal domain have grown rapidly, yielding promising results in tasks such as contract analysis, judgment summarization, legal consultation, and case prediction . To promote research in legal language processing, several benchmark datasets and evaluation platforms for LLMs have emerged in recent years, including LexGLUE (Chalkidis et al., 2021, 2020), LegalBench (Guha et al., 2023), and the COLIEE competition on statutory entailment and retrieval. These benchmarks primarily cover tasks such as multiple-choice questions, case classification, statute matching, and legal question answering. However, most of them focus on English-language corpora and closed-form problems, lack-

* Corresponding author. ORCID: 0000-0002-7101-5272

ing the design needed to evaluate the type of open-ended, reasoning-intensive essay questions encountered in real-world legal practice. As noted by the creators of LegalBench, current benchmarks "still fall short of comprehensively evaluating the open-ended reasoning required in law school exams and legal writing assignments" (Guha et al., 2023), which often involve deep statutory subsumption and integrated legal analysis.

Chain-of-Thought (CoT) prompting has recently emerged as a promising strategy for improving multi-step reasoning and computation, validated on tasks such as math word problems and commonsense reasoning (Wei et al., 2022). CoT prompting has to be effective even without in-context examples: simple natural language cues like "Let's think step by step" can activate internal reasoning chains and significantly improve performance in zero-shot settings (Kojima et al., 2022). CoT has since been widely applied in mathematical reasoning (e.g., GSM8K, MATH), logic puzzles, scientific domains, and programming tasks. Prior studies consistently find CoT especially useful for tasks that require intermediate inference steps, as it helps maintain contextual coherence and supports longer, structured chains of reasoning.

Although legal reasoning itself is inherently a multi-step logical task, systematic analysis of CoT prompting in the legal domain remains limited. Some notable attempts include the KIS team's Interpretable CoT strategy in the COLIEE 2024 entailment task, which enhances interpretability in statutory subsumption through structured prompting (Fujita et al., 2024). Another example is the LegalGPT framework, which integrates CoT modules within a multi-agent architecture to simulate the collaborative logic of real-world legal practice (Shi et al., 2024).

Mainstream approaches to evaluating legal LLMs typically rely on automated metrics (e.g., accuracy, BLEU, F1-score) or multiple-choice style datasets to compare model performance. However, such closed-form evaluations fail to reflect the logical depth and reasoning quality required for open-ended generative tasks. Recent studies have begun to incorporate human evaluation to better assess consistency and subsumption performance in long-form legal QA. Representative systems such as Length-Controlled AlpacaEval (Dubois et al., 2024), MT-Bench (Zheng et al., 2023) , and PromptBench (Jiang et al., 2023) use human preference ratings or expert judgments as quality signals, augmented by ranking-based metrics, weighted averages, or Elo-style comparisons.

we conduct qualitative analysis by selecting cases with high inter-rater agreement to compare the logical structure of reasoning chains, thereby supplementing the current lack of process transparency and error traceability in legal LLM evaluation.

3 Experiment Design

3.1 Stage 1: Decomposed Reasoning Evaluation

3.1.1 Test Set Design

The test set used in this study consists of six essay questions, all adapted from previous Judicial Officer Examinations and the National Taiwan University Graduate Law School entrance exams in the field of Status Law. These questions cover a range of key topics, including the validity of marriage, division of marital property, legal guardianship, inheritance, bigamy, and adoption.

The test set comprises six essay questions adapted from Taiwan's judicial exams and law school admissions, focusing on key topics in Status Law such as marriage validity, inheritance, and guardianship. Each question was decomposed into multiple sub-questions—68 in total—targeting factual analysis, statutory application, and logic chaining. Status Law was selected due to its blended demands of symbolic reasoning, legal interpretation, and numerical computation, making it an ideal domain for evaluating LLMs' integrated legal reasoning capabilities.

The six essay questions selected for this study cover the following legal topics: **A. Validity of marriage, B. Division of residual marital property, C. Limitations on parental rights in relation to children's interests, D. Limited succession and creditor claims, E, Legal consequences of bigamy, F. Collation issues in inheritance distribution.**

The sub-questions are primarily framed as numerical problems and binary (yes/no) questions, with a few short-answer questions. Each sub-question presents a specific legal scenario and requires the LLM to provide a definitive judgment or calculation. To ensure stricter evaluation, this study adopts a rigorous scoring standard: if the model arrives at the correct final answer but misidentifies roles, relationships, or inheritance rankings within its reasoning, the response is marked incor-

rect. This prevents models from "guessing correctly" and emphasizes the need for accurate legal reasoning and comprehension.

3.1.2 Model Selection

The study evaluates four mainstream LLMs: ChatGPT-4o (OpenAI), Grok 3 (xAI), Gemini 1.5 Flash (Google), and Copilot (Microsoft, based on GPT-4-turbo). These models were selected to represent the current state-of-the-art offerings from the four major LLM platforms, balancing accessibility, popularity, and architectural diversity. All models were tested under identical formats and prompt templates to ensure fairness.

While some LLMs—such as ChatGPT or Gemini—exhibit emergent Chain-of-Thought (CoT) reasoning capabilities without explicit prompting, our experiments show that these models still often display under-reasoning behaviors, skipping intermediate legal logic steps or prematurely concluding without sufficient statutory justification.

Rather than designing a universal prompt for all models or modifying LLM parameters, our approach centers on a specialist-in-the-loop interaction. Here, a legal expert provides real-time validation at each reasoning stage, enabling targeted scaffolding and reducing cumulative error. This controlled feedback setup aims not merely to improve answer correctness, but to systematically expose and align the model's legal reasoning flow with doctrinal logic. Such design reflects pedagogical principles of legal education and offers a practical pathway for integrating LLMs into high-stakes, explainable legal workflows.

These sub-questions were manually constructed to reflect three core legal reasoning types: factual classification, statutory mapping, and inheritance computation. Each sub-task is framed as either a binary (yes/no), numerical, or short-form reasoning prompt. This design captures both surface-level legal recognition and deeper multi-step logic.

3.2 Stage Two: Comparison of CoT vs. Direct Answering

3.2.1 Response Format and Prompting Design

In Stage Two, models were tested on the full legal questions under two conditions: baseline answering (one-shot response) and Chain-of-Thought (CoT) prompting.

Responses were evaluated by a professor and a law student, blind to the model type and prompting condition. Scoring followed a 0–10 rubric covering issue coverage, statutory accuracy, and reasoning clarity. This design simulated realistic exam settings while enabling structured analysis of model reasoning quality.

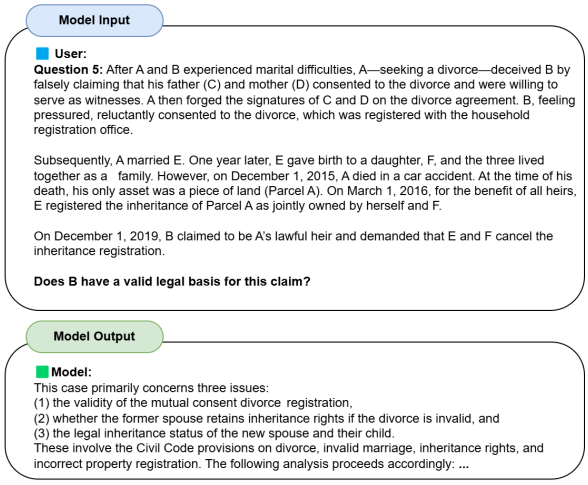

Figure 1: Example of Baseline Prompting

In contrast, Figure 2 depicts the Chain-of-Thought (CoT) prompting setting. The model is prompted to answer each sub-question step by step, with immediate user feedback at each stage to confirm or correct the response. This interactive scaffolding prevents early-stage reasoning errors from propagating and encourages structured legal analysis. After completing all sub-steps, the model is prompted to synthesize a full-length answer based on the verified intermediate results.

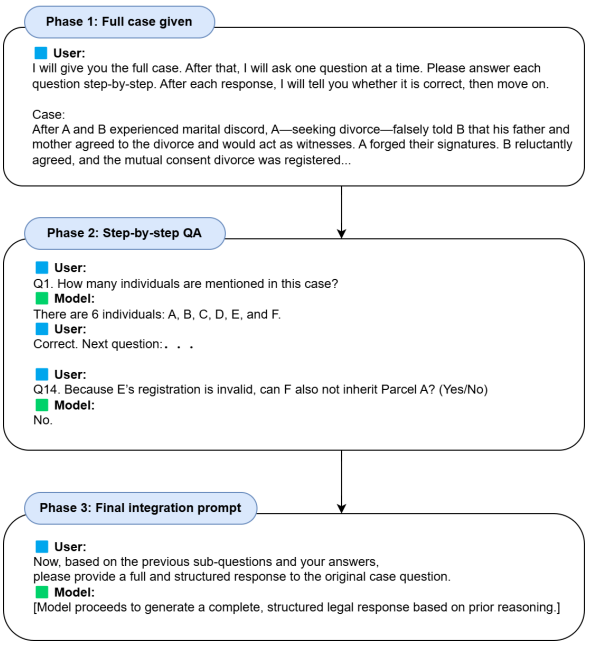

Figure 2: Example of CoT Prompting

This contrast illustrates how CoT prompting transforms the model's reasoning process from a monolithic, opaque response into an interactive, modular sequence of logic steps, allowing for clearer observation and evaluation.

3.2.2 Scoring Mechanism

Both response versions were evaluated by scorers with formal legal training, who rated each answer holistically on a 0–10 scale, with higher scores reflecting better overall quality. Scoring was based on three key criteria:

1. Issue coverage: Whether the model identified and addressed the key legal issues and factual disputes in the question.

2. Accuracy of statutory application: Whether the cited or applied legal provisions were correct and logically relevant.

3. Clarity of legal reasoning: Whether the reasoning was coherent, structured, and logically sound.

4 Experiment Result

4.1 Stage One Results: Accuracy in Decomposed Reasoning Evaluation

This stage also examined the effect of Chain-of-Thought (CoT) prompting on answer accuracy. The table below presents the accuracy performance of the four LLMs under baseline and CoT conditions, along with results from a paired-sample t-test:

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

Table 1: Accuracy comparison between raw and CoT prompting across LLMs

Overall, even though the unit of observation in this stage was the model's performance on decomposed sub-questions, the results clearly demonstrate that CoT prompting significantly enhanced accuracy for certain models. The effect was particularly pronounced in question types that involved multi-step reasoning and structured analysis, such as inheritance calculation, classification of legal status relationships, and precise statutory mapping. These findings provide a quantitative foundation for the holistic answer evaluations conducted in Stage Two.

4.2 Stage Two Results: Human Evaluation of Full-Length Responses Under CoT Prompting

The goal of the Stage Two experiment was to simulate realistic legal exam conditions and assess whether the overall quality of LLM-generated responses to unsegmented, full-length legal questions could be improved by introducing Chain-of-Thought (CoT) prompting. The same six Status Law questions used in Stage One were employed, but this time they were presented in their entirety—without decomposition—requiring the model to generate a complete answer in one go.

Model	**Raw** *Ave. Score*	**CoT** *Ave. Score*	**Average Improv.**	**Improv. Stud.** *Rater*	**Improv. Prof.** *Rater*	**Scoring Consistency** *(Pearson's r)*
ChatGPT	6.50	9.17	**+2.67**	+3.00	+2.33	0.716
Gemini	6.12	8.04	+1.92	+2.83	+1.00	0.853
Copilot	5.83	7.42	+1.58	+2.00	+1.17	0.752
Grok3	6.25	8.08	+1.83	+2.17	+1.50	0.835

Table 2: Human evaluation results under baseline vs. CoT prompting

4.2.1 Overall Model Scoring Results

The results show that all models demonstrated improved performance when CoT prompting was applied. Among them, ChatGPT exhibited the largest improvement (+2.67 points) and the most consistent performance across questions. Gemini and Grok3 also showed marked improvements, each with gains exceeding 1.8 points. Although Copilot lagged behind the other models in terms of raw scores, it too displayed consistent improvement under CoT prompting.

In terms of inter-rater agreement, Pearson correlation coefficients ranged from 0.71 to 0.85 across the four models, indicating a moderate to high level of scoring consistency between the two raters. Notably, Gemini and Grok3 achieved the highest consistency, suggesting particularly stable performance as evaluated by both expert and student raters.

Q.	Raw Avg.	CoT Avg.	Score Gain	p-value
1	4.50	7.75	+3.25	0.068
2	6.12	7.12	+1.00	0.430
3	6.25	6.38	+0.12	0.919
4	3.88	5.50	+1.62	0.080
5	3.88	7.00	+3.12	0.002
6	4.63	7.50	+2.88	0.011

Table 3: Average scores for each legal question under raw vs. CoT prompting

4.3 Qualitative Analysis of Model Responses on a Representative Question

To illustrate model reasoning differences, we selected Question 5 as a representative case based on high inter-rater agreement. Gemini's baseline response exhibited flawed assumptions and incorrect citations, such as misapplying Article 92 instead of the correct divorce statute. Its CoT-prompted version showed clearer structure and partial legal improvement, yet still missed key statutes and overgeneralized inheritance logic. This contrast highlights CoT's benefit in structuring legal reasoning, though gaps remain in precise statutory application and doctrinal subsumption (see Appendix A for question details).

Table 4 provides a visual comparison of key statutes that should be cited in an ideal answer like Figure 3.

Issue	Original Reasoning	CoT Reasoning
Divorce	Assumes validity; cites §92 (intent defect)	Finds formal defect; cites §1050
Remarriage	Treats E as lawful spouse	Void due to bigamy; E not in good faith
F's Inheritance	Assumes F is legitimate heir	F is non-marital but legally recognized
Registration	Both E and F must cancel	Only E is void; F retains right

Table 4: Comparison of original and CoT reasoning on four legal issues.

The original version erroneously cited Article 92, which pertains to the revocation of declarations of intent due to fraud. This reflects a misunderstanding of the legal nature of the problem—it misclassified the issue as a defect in intent rather than a formal defect that invalidates the divorce. Furthermore, it failed to mention several key statutes, including those governing bigamy and non-marital inheritance.

The CoT version correctly cited Articles 1050 and 1138, capturing part of the statutory logic. However, it omitted Article 767 and did not clearly reference the provisions governing non-marital children, resulting in a fragmented presentation of the statutory framework.

5 Conclusion and Limitations

5.1 CoT-on-CoT

This paper presents a two-stage diagnostic framework to evaluate how Large Language Models

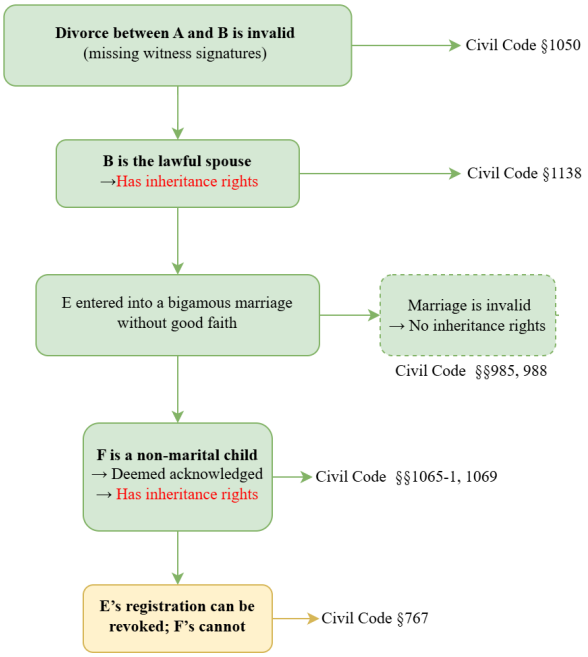

Figure 3: Correct legal reasoning flowchart

(LLMs) reason through legal essay questions in Taiwan's Status Law. By decomposing real exam questions into 68 sub-tasks and comparing full-length responses under baseline and Chain-of-Thought (CoT) prompting, we assess both micro-level reasoning and holistic legal understanding. Results show that CoT significantly improves accuracy in issue spotting, statutory application, and inheritance calculation, particularly for Gemini and Grok3.

Human evaluation further reveals enhanced clarity and structure in CoT-generated answers, though alignment gaps with expert persist. While our focus is not on prompt universalization, but explore CoT-on-CoT designs using models already trained with internal reasoning strategies, to examine whether reasoning stability or redundancy effects emerge. The CoT prompting can effectively improve the logical structure and issue coverage in model-generated responses.

Rather than fully opening the black box of LLM reasoning, our approach traces the model's internal chains of thought by eliciting and examining intermediate steps, thereby making its legal decision path more interpretable. We release a legally grounded benchmark and propose a generalizable evaluation methodology for open-ended, multi-step reasoning tasks.

5.2 Limitations and Future Work

This study has several limitations that open avenues for future work. First, while our results demonstrate that LLMs often produce incorrect reasoning even when their final answers are right, we do not yet offer a systematic typology of such reasoning failures. Future research could develop finer-grained error categories and explore how these missteps relate to different legal domains or prompt structures.

Second, although our methodology involves decomposing legal questions into sub-tasks, we have not formalized a reusable guideline for annotators or model developers to construct reasoning flowcharts. Creating such a protocol—potentially in the form of annotation templates or instructional schemas—could support replicability and improve human-LLM alignment in legal diagnostics.

Third, while we propose three evaluation dimensions (issue coverage, statutory application, reasoning clarity), we have not validated their generalizability across legal domains beyond Taiwanese Status Law. We believe these dimensions are transferable, but further experiments on multilingual or cross-jurisdictional datasets (e.g., U.S. torts, Japanese family law) are needed to assess the framework's robustness and scalability.

Finally, our current evaluation focuses on a limited set of essay-style questions, and has not been tested at scale. Integration with existing legal benchmarks (e.g., LegalBench, COLIEE) could allow broader adoption, while future work may automate sub-question generation or integrate CoT supervision into fine-tuning pipelines.

References

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*.

Yuan Chen, Ronglai Shen, Xiwen Feng, and Katherine Panageas. 2024. Unlocking the power of multi-institutional data: Integrating and harmonizing genomic data across institutions. *Biometrics*, 80(4):ujae146.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Masaki Fujita, Takaaki Onaga, and Yoshinobu Kano. 2024. Llm tuning and interpretable cot: Kis team in coliee 2024. In *JSAI International Symposium on Artificial Intelligence*, pages 140–155. Springer.

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.

Yue Jiang, Siyu Zheng, Xin Chen, Hao Peng, Xiang Lin, and 1 others. 2023. Promptbench: Towards evaluating the robustness of large language models with prompt-based benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Juanming Shi, Qinglang Guo, Yong Liao, and Shenglin Liang. 2024. Legalgpt: Legal chain of thought for the legal large language model multi-agent framework. In *International Conference on Intelligent Computing*, pages 25–37. Springer.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Siyu Zheng, Bo Peng, Hao Peng, Zhen Zhang, Yao Shen, Zhuohan Li, Weifeng Du, Tao Yu, Tianyi Zhang, Xiang Lin, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Appendix: Test Set and Sub-question Decomposition

Design Rationale. The test set includes six legal essay questions adapted from Taiwan's Judicial Officer Examinations, National Taiwan University Law Graduate Admissions Exams, and final assessments in Identity Law. Each question targets core legal topics and is decomposed into sub-questions assessing factual comprehension, legal classification, statutory application, and logical reasoning.

Question: Bigamy and Inheritance Disputes

Question A and B registered a consensual divorce based on forged witness signatures. A later married E and had a daughter F. A died in a car crash, and E registered inheritance jointly with F. B later claims as the legal heir. Is her claim valid?

Sub-questions

1. How many individuals are involved?
2. Were A and B legally married? (Yes/No)
3. How many times did A marry?
4. Is divorce valid without witness verification? (Yes/No)
5. Was A and B's divorce legally valid? (Yes/No)
6. If A never divorced B, is marriage to E valid? (Yes/No)
7. If E was unaware, does good faith validate marriage? (Yes/No)
8. Is B a legal heir? (Yes/No)
9. Is E a legal heir? (Yes/No)
10. Are C and D (A's parents) legal heirs? (Yes/No)
11. Is F excluded due to being non-marital? (Yes/No)
12. Does E's co-ownership registration remain valid? (Yes/No)
13. If E is not a legal heir, is her registration invalid? (Yes/No)
14. If E's registration is invalid, does it affect F's inheritance? (Yes/No)