

# CIS-RG at QIAS 2025 Shared Task: Chain of Thought Prompting and Finetuning for Enhancing Performance of LLMs on Islamic Legal Reasoning and its Mathematical Calculations

Osama Zaki<sup>1,2</sup>, Asmaa Badawy<sup>2</sup>, Nada Elgewily<sup>2</sup>, Ahmed Sharaf<sup>2</sup>

1. ROBOTAAAR, UK

2. Sinai University, El-Arish, Egypt

## Abstract

The work in this paper is related to the shared task QIAS2025. In this paper we continue in assessing large language models on Islamic legal reasoning. It is a challenging task because LLMs have not yet evolved (especially the open-source models) to solve complex reasoning problems or to perform mathematical calculations that require several steps. The LLMs need to comprehend the problem and to generate accurate and justified answers. In this paper we confirm the results and the analysis given in (Boucekif, 2024; Boucekif, 2025). However, we experiment further with Fine Tuning and Chain of Thought (CoT) to improve the performance of the reasoning process and therefor the results of the LLMs.

## 1 Introduction

This shared task assesses the ability of LLMs to accurately answer questions about ‘Im al-mawārīth (The science of Islamic Inheritance) in realistic scenarios. It is a major specialized topic in Islamic law. Islamic inheritance’s rules are well defined, but it requires a complex reasoning mechanism and well-designed and systematic calculations procedures. There are mainly three computational stages, each includes zero, one or more than one step, that required to solve an Islamic inheritance case. First stage is to comprehend the inheritance scenario presented to

the system, to identify eligible and the non-eligible heirs based on their relationship to the deceased person, bequests, the distribution of a defined amount of money, blocking or exclusion of some heirs, and to apply the basic fixed-share rules (farā’id). Second stage is to consider the cases where there are multiple heirs, multiple deceased individuals, residuary shares, and partial exclusion. Third stage is to consider the intricate fractional calculations, adjusting and redistribution, exaptational and nuanced cases, and juristic disputes. Although those stages seems like they can be carried out in sequence they are Intertwined and the system it has go back and forward over the rules. This makes the science of inheritance complex due to its diverse situations, the multiplicity of heirs, and the factors affecting the calculation of the estate, which requires a precise understanding of the texts of Islamic law and their correct application to prevent disputes and achieve justice in the distribution of rights.

## 2 System Design Issues

In (Boucekif, 2024; Boucekif, 2025) the performance of seven LLMs were assessed using a benchmark of 1,000 multiple-choice questions covering diverse inheritance scenarios, designed to test each model’s ability to solve such problems. Gemini 2.5 and o3, demonstrated high performance, achieving accuracy above 90%. GPT4.5 achieved moderate results. Jais, Mistral, and LLaMA showed significantly lower

accuracy reflecting their limitations in legal reasoning. There is a clear gap between models with reasoning abilities and those without. ALLaM, Fanar, LLaMA, and Mistral, consistently struggled with identifying complex familial relationships, evaluating diverse inheritance scenarios, and correctly executing corrective calculations.

As shown in the following section we assessed four models: Fanar, Llama, Gemini and Mistral. The models are further fine-tuned with a well-defined and large set of 1000 examples. We also recognized that the model architecture plays a major role in the result, i.e., being capable of performing reasoning or not. Models with reasoning capabilities consistently perform better. Having stated that, the reasoning capability is usually built outside the core of the model it is usually build at the application layer, i.e. the prompt, being the layer representing both the input and the output of the model. Many models nowadays claim that they have reasoning capability or at least able to respond correctly to simple reasoning task, but the challenge however among models present when dealing with complex reasoning problems.

LLMs evolved from just being a next-token prediction task dealing mainly with natural language (Zhao, 2023), to code generation (Gehring, 2024), and logical reasoning (Webb, 2023). Techniques such as chain of thoughts prompting techniques (Wei, 2022), tree of thought (Yao, 2024), trial-and-error search (Luo, 2024), Process Reward Models which facilitate reinforcement learning for LLMs (Sun, 2024). These emergent techniques are based on two main concepts in the traditional AI: “search” and “learning”. A combination of scaling train-time compute and test-time-compute leads to better reasoning performance (OpenAI, 2024). To sum up, there is main four approaches for reasoning: 1) chain-of-thought (CoT) prompting which increases computational resources during inference to improve output quality. 2) Pure reinforcement learning (RL) 3) supervised finetuning (SFT) 4) combining both RL and SFT (Raschka, 2025).

In this paper, CoT prompting combined with finetuning is being the focus in our investigation. To do that, different finetuning datasets were prepared with different sizes (e.g., 100 and 200 questions) representing two different clusters. The

first contains samples without any mathematical calculations, while the second contains samples that require mathematical calculations. As an example for the first cluster:

مات وترك: عم الأب لأب (4) وأخت لأب (5) و عم لأب (2) و أم أم الأب وأم أم الأم كم النصيب الأصلي لـ عم لأب (2) من التركة، وما الدليل على ذلك؟

The chain of thought (step-by-step) that should be followed is:

1. The type of actors in the question:  
وأم أم الأم، وأم أم الأب، وعم لأب، وأخت لأب، وعم الأب لأب
2. Those who deserve a fixed share:  
وأم أم الأم، وأخت لأب، وأخت لأب
3. Those who deserve a non-fixed share:  
عم لأب
4. Those who are blocked:  
عم الأب لأب
5. From the above,  
عم لأب هو من العصبات بالنفس، ويرث ما بقي من التركة بعد أصحاب الفروض
6. The number of actors in each type  
ما بقي من التركة يقسم على اثنين

To list the steps for each question (case) in this manner is unrealistic, but it is possible only for few shots. However, it still requires considerable efforts and skills to integrate CoT with the MCQs dataset.

Results of CoT promoting approach is still under investigation. In the following section we analysis the results of the traditional finetuning approach without the implementation of CoT.

### 3 Experimental Setup, Results and Analysis:

Four LLMs were fine-tuned: Fanar, Llama, Gemini and Mistral. The models were tested on the provided dataset by the shared task which contains unlabeled 1000 MCQs questions (answers is one of the letters: A, B, C, D, E or F, i.e. six choices).

The results are Gemini 2.5 and o3, demonstrated high performance, achieving accuracy above 90%. Fanar, Mistral, and LLaMA achieved moderate results 76%, 74%, 73% respectively reflecting their limitations in legal reasoning. This confirms to the findings in

Four examples form the test dataset are selected to demonstrate different scenarios, Table [1]. Our

analysis shows that the level of the question being beginner or advanced has some impact on the model, but it is the wording of the question is the main reason that makes the model comprehend the questions. This means that classifying the questions to beginner and advanced are not very useful.

Question ID	Gold	Level	Fanar	Llama	Gemini	Mistral
8804_n11d9s7s_4	E	Beginner	E	C	E	E
7398_nf5b3z3e_11	A	Advanced	D	C	A	C
8545_ni6z316j_12	D	Advanced	C	F	D	F
6839_np2c3f4i_4	B	Beginner	A	D	A	F

Table 1: Results from Fanar, Llama, Gemini and Mistral

But rather we can classify the questions into logical and mathematical problems. It is obvious that the commercial models are better pre-trained and hence fine-tuned. The generated justification and calculations of the commercial models were better represented and more detailed. The four selected example are (Fanar and Gemini only):

#### Question 8804\_n11d9s7s\_4

مات وترك: عم الأب لأب (4) و أخت لأب (5) و عم لأب (2) و أم أم الأب وأم أم الأم كم النصيب الأصلي لـ عم لأب (2) من التركة، وما الدليل على ذلك؟

"نصيبه هو لا شيء، والدليل: لأنه عصبية"، "نصيبه هو باقي التركة، والدليل: لأنه عصبية"، "نصيبه هو محجوب، والدليل: لأنه عصبية"، "نصيبه هو كل التركة، والدليل: لأنه عصبية"، "نصيبه هو باقي التركة، والدليل: لأنه عصبية"، "نصيبه هو النصف، والدليل: لأنه عصبية"

Gemini

التعليل: عم لأب هو من العصبات بالنفس، ويرث ما بقي من التركة بعد أصحاب الفروض (الأخت لأب والجدة). وقد أجمع العلماء على أن العصبات يرثون ما فضل عن نصيب أصحاب الفروض، عملاً بقول النبي صلى الله عليه وسلم: "أحقوا الفرائض بأهلها، فما بقي فلأولى رجل ذكر". وهو هنا أقرب العصبات الموجودين في المسألة (أقرب من عم الأب لأب) فيأخذ الباقي.

Fanar

التعليل: لأن العم لأب من العصبات، وعند وجود عصبية من الذكور، يأخذون ما تبقى بعد الفرائض، ولا يوجد في هذه المسألة ذكور عصبية غير العم لأب، لذا فهو يأخذ الباقي من التركة. (ولكن يجب التنويه أن العم لأب ليس أولى العصبات، بل يأتي بعد الأب والجدة وأخو الأب وأم الأب في الترتيب، ولكن في هذه المسألة، لا يوجد من هو أولى منه من الذكور العصبية، لذا فهو يأخذ الباقي)

Three models were able to produce the correct answer. The Justifications given by Fanar and Gemini are also correct, but Gemini style is clearer. Although a fault was injected into the

question (e.g., choices B and E), both exactly have the same wording, the three models have not detected that and give the answer as E.

#### Question 7398\_nf5b3z3e\_11

مات وترك: أم وأخ لأم (3) و ابن ابن أخ لأب (2) و ابن ابن (2) و ابن ابن أخ شقيق وأخ شقيق (4)، كم إجمالي عدد الأسهم الذي تقسم عليه التركة قبل تصحيح المسألة؟ 6، 7، 8، 4، 5، 0 أسهم

Gemini

التعليل: الأم ترث السدس لوجود الفرع الوارث (ابن الابن)، لقوله تعالى: "ولأبويه لكل واحد منهما السدس مما ترك إن كان له ولد". وابن الابن يحجب جميع الإخوة (الأخ الشقيق، والأخ لأب، والأخ لأم) وأولادهم. وبناءً عليه، يكون أصل المسألة من مقام نصيب الأم وهو 6 أسهم.

Fanar

التعليل: يبدأ حساب الأسهم من أصل المسألة الذي يخرج منه فرض الأم، وهو ثمانية أسهم، حيث تأخذ الأم الثمن (واحد) والباقي للفرع الوارث تعصيباً

Only Gemini was able to generate the correct answer for this advanced question. The justification given by Gemini is correct and it is detailed, while justification from Fanar is incorrect. This shows that Fanar didn't comprehend the problem by saying:

وهو ثمانية أسهم، حيث تأخذ الأم الثمن (واحد) والباقي للفرع الوارث تعصيباً

#### Question 8545\_ni6z316j\_12

مات وترك: أب الأب و أخ لأم (3) و أخ شقيق (2)، كم عدد الأسهم بعد التصحيح التي يحصل عليها لكل أخ شقيق؟ 4، 5، 0 أسهم، 1 سهم، 3 أسهم، سهمان

Gemini

التعليل: في هذه المسألة، يحجب أب الأب (الجد الصحيح) الإخوة لأم بالإجماع. أما بالنسبة للإخوة الأشقاء مع الجد، فإن مذهب الجمهور يقضي بأن الجد يقاسم الإخوة أو يأخذ ثلث المال (أيهما أفضل له). في هذه الحالة، لا يوجد أصحاب فروض آخرون، فالجد يقاسم الإخوة الأشقاء (وعدد هم 2).

إذا اعتبرنا أصل المسألة 3 سهام:  
\* الجد يأخذ 1 سهم (3/1 من التركة).  
\* الأخوان الشقيقان يأخذان 2 سهم (3/2 من التركة).  
\* نصيب كل أخ شقيق = 2 سهم / 2 أخ = 1 سهم.

Fanar

التعليل: لا يرث الأخ لأم مع وجود الأخ الشقيق، فالأخ الشقيق يحجب الأخ لأب.

Gemini was able to generate the correct answer for this advanced question. The justification and the explanation of mathematical steps given by Gemini are correct and clear which includes

fractional allocation, while the justification from Fanar is incorrect which means Fanar has not understood this case (question).

#### Question 6839 np2c3f4i 4

مات وترك: عم الأب لأب و أم أم الأم و ابن ابن عم شقيق (3) و زوجة كم النصيب الأصلي لـ ابن ابن عم شقيق (3) من التركة، وما الدليل على ذلك؟ "

نصيبه هو باقي التركة ، والدليل: ترتيب العصابات يكون أولاً بالجهة ثم بالدرجة ثم بالقوة فمن حيث الجهة يقدم الأبناء الذكور على الأصول -الأب وأب الأب وإن علا - ، ومن حيث الدرجة يقدم من علا على من نزل فيقدم الإبن على الأب والإبن وأب على أب الأب ، ومن حيث القوة يقدم من كانت قرابته بأبوين على من كان بأب فقط فيقدم الأخ الشقيق على الذي لأب ويقدم ابن الأخ الشقيق على ابن الأخ لأب ، فلا يرث أبعد بتعصيب مع أقرب منه ، لقوله ﷺ (الحقوا الفرائض بأهلها فما بقي فلأولى رجل ذكر). وهنا أولى رجل ذكر هو ابن ابن عم شقيق، لذا يكون نصيبه باقي التركة .

رجل ذكر هو ابن ابن عم شقيق" ،  
" نصيبه هو باقي التركة، والدليل: ترتيب العصابات يكون..  
" نصيبه هو باقي التركة، والدليل: ترتيب العصابات يكون...  
" نصيبه هو باقي التركة، والدليل: ترتيب العصابات يكون...  
" نصيبه هو باقي التركة، والدليل: ترتيب العصابات يكون...  
" نصيبه هو باقي التركة، والدليل: ترتيب العصابات يكون

#### Gemini

تعليل: في هذه المسألة، لدينا أصحاب فروض وعصابات. الزوجة: ترث الربع (4/1) لعدم وجود فرع وارث للميت، لقوله تعالى: "وَلَهُنَّ الرُّبُعُ مِمَّا تَرَكْتُمْ إِنْ لَمْ يَكُنْ لَكُمْ وَاوْلَادٌ" (النساء: 12).

أم أم الأم: هي جدة صحيحة، وترث السدس (6/1) عند عدم الأم، وهو قول الجمهور.  
ابن ابن عم شقيق (3) وعم الأب لأب: كلاهما من العصابات بالنفس. يتم تقديم العصابات بالترتيب: الجهة، ثم الدرجة ثم القوة.  
كلاهما من جهة العمومة.

بين ابن ابن عم شقيق وعم الأب لأب، ابن ابن العم الشقيق أقرب في الدرجة والقوة. ابن ابن العم الشقيق هو من فروع أعمام الميت، بينما عم الأب لأب هو من أعمام أب الميت (أي أعمام الجد). الأقرب من جهة الميت يقدم، ففروع أعمام الميت مقدمون على أعمام الأب.

لذلك، يجب عم الأب لأب بابن ابن عم شقيق.  
يأخذ ابن ابن عم شقيق (3) باقي التركة بعد أصحاب الفروض.  
الحساب:

أصل المسألة من 12 (المضاعف المشترك الأصغر للمقامات 4 و 6).

نصيب الزوجة:  $12 \div 4 = 3$  أسهم.  
نصيب أم أم الأم:  $12 \div 6 = 2$  سهمين.  
مجموع أسهم أصحاب الفروض:  $3 + 2 = 5$  أسهم.  
الباقي للعصابة (ابن ابن عم شقيق (3)):  $12 - 5 = 7$  أسهم.

الدليل العام للعصابات هو قوله ﷺ: "ألقوا الفرائض بأهلها فما بقي فهو لأولى رجل ذكر" (صحيح البخاري ومسلم).  
وهنا أولى رجل ذكر هو ابن ابن عم شقيق.

#### Fanar

التعليل: ترتيب العصابات يكون أولاً بالجهة، ثم بالدرجة ثم بالقوة، فمن حيث الجهة يقدم الأبناء الذكور على الأصول (الأب وأب الأب وإن علا)، ومن حيث الدرجة يقدم من علا على من نزل فيقدم الإبن على ابن الإبن والأب على أب الأب، ومن حيث القوة يقدم من كانت قرابته بأبوين على من كان بأب فقط فيقدم الأخ الشقيق على الذي لأب ويقدم ابن الأخ الشقيق على ابن الأخ لأب، فلا يرث أبعد بتعصيب مع أقرب منه، لقوله ﷺ (الحقوا الفرائض بأهلها فما بقي فلأولى رجل ذكر). وهنا أولى رجل ذكر هو ابن ابن عم شقيق، لذا يكون نصيبه باقي التركة .

This example demonstrates a basic question, but its prompt has some issues. The evidence referred to by each choice is identical for all of them. This

#### 4 Conclusion and Future Work

Our experiments with finetuning have revealed some important points: 1) the classifications of questions into two or three categories (intermediate, advanced) have not shown noticeable impact on the results, 2) the uncleanes of the training or the validation dataset has some impact on the results, 3) LLMs with no reasoning capabilities (mostly the open source LLMs) struggle to solve complex reasoning problems, 4) MCQ is not the optimal option to train, validate especially when representing an 'evidence' with the 'choice', and this evidence is shared among other choices.

However, our initial experiments (ongoing work) with CoT have shown some promising results. We plan to combine CoT with RL. We intend classifying the questions (datasets) into two clusters: logical thinking problems and mathematical calculations problems.

We also believe a hybrid approach, agentic AI or neuro-symbolic systems, which can reason step by step, in algorithmic manner, that adhere exactly and precisely to legal rules and adapt to complex inheritance cases will enhance the performance.

Finally, when dealing with legal and/or religious domain such as the Islamic inheritance the LLMs responses should be verified by a legal lawyer or in a court.

## References

- Abdessalam Boucekif, Samer Rashwani, Heba Sbahi, Shahd Gaben, Mutaz Al-Khatib, and Mohammed Ghaly. 2025. *Assessing Large Language Models on Islamic Legal Reasoning: Evidence from Inheritance Law Evaluation*, Proceedings of The Third Arabic Natural Language Processing Conference, Suzhou, China. Association for Computational Linguistics.
- Abdessalam Boucekif and Samer Rashwani and Emad Mohamed and Mutaz Al-Khatib and Heba Sbahi and Shahd Gaben and Wajdi Zaghouni and Aiman Erbad and Mohammed Ghaly. 2025. *QIAS 2025: Overview of the Shared Task on Islamic Inheritance Reasoning and Knowledge Assessment*. Proceedings of The Third Arabic Natural Language Processing Conference, ArabicNLP 2025, Suzhou, China, November 5--9, 2025. Association for Computational Linguistics.
- Jonas Gehring, Kunhao Zheng, Jade Copet, Vegard Mella, Taco Cohen, and Gabriel Syn-naeve. 2024. Rlef: Grounding code llms in execution feedback with reinforcement learning. arXiv preprint arXiv:2410.02089.
- Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, et al. 2024. Improve mathematical reasoning in language models by automated process supervision. arXiv preprint arXiv:2406.06592.
- Sebastian Raschka. 2025. Understanding Reasoning LLMs. Ahead AI. <https://magazine.sebastianraschka.com/p/understanding-reasoning-llms>
- Chuanneng Sun, Songjun Huang, and Dario Pompili. 2024. Retrieval-augmented hierarchical in-context reinforcement learning and hindsight modular reflections for task planning with llms.2024.
- OpenAI. 2024. Learning to reason with llms.
- Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.
- Osama Zaki. 2024. Coupling Machine Learning with Ontology for Robotics Applications. *Robotics*. arXiv:2407.02500.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. arXiv preprint arXiv:2303.18223.