Towards Grounded Legal AI

Fabio J. Fehr

OXFORD ARTIFICIAL
INTELLIGENCE SOCIETY

Rilton Franzone*

Valentin Noël*

Puyu Wang

Philip Torr

Fabio J. Fehr

UNIVERSITY OF OXFORD

OXFORD ARTIFICIAL INTELLIGENCE SOCIETY

midpage

devoteam

Building Reward Models for Grounded Legal Reasoning

Rilton Franzone [*1] **Valentin Noël** [*2] **Puyu Wang** [3 4] **Philip Torr** [3] **Fabio J. Fehr** [3 4]

Benchmark Code

Abstract

Large language models are increasingly used in high-stakes domains such as law, where systems must ground their reasoning in evidence and abstain under uncertainty. However, existing reward models prioritise helpfulness and fluency rather than contextual grounding, often producing unsupported legal reasoning in retrieval-augmented generation (RAG) settings. We introduce LEGAL-REWARDBENCH (LRB), a legal contextual reward modelling benchmark constructed from legal QA datasets for evaluating grounded legal

1. Introduction

Large language models (LLMs) are increasingly deployed in high-stakes domains such as law, medicine, and finance, where generated outputs may influence legal decisions, clinical care, and access to essential services (Siino et al., 2025; Costa-Gomes et al., 2026; Chen et al., 2024). Unlike open-domain conversational settings, these domains require reasoning grounded in external evidence and constrained by domain-specific rules. In such settings, useful systems must not only produce correct answers, but also remain faithful to the provided context and refuse to answer when the available evidence is insufficient (Pipitone & Alami, 2024; Peng et al., 2025). However, current language models are primarily optimised for fluent and helpful interaction, often

ICML
International Conference
On Machine Learning

ACCEPTED!

Building Reward Models for Grounded Legal Reasoning

Rilton Franzone [*1] Valentin Noël [*2] Puyu Wang [3 4] Philip Torr [3] Fabio J. Fehr [3 4]

Benchmark Code

Abstract

Large language models are increasingly used in high-stakes domains such as law, where systems must ground their reasoning in evidence and abstain under uncertainty. However, existing reward models prioritise helpfulness and fluency rather than contextual grounding, often producing unsupported legal reasoning in retrieval-augmented generation (RAG) settings. We introduce LEGAL-REWARDBENCH (LRB), a legal contextual reward modelling benchmark constructed from legal QA datasets for evaluating grounded legal

1. Introduction

Large language models (LLMs) are increasingly deployed in high-stakes domains such as law, medicine, and finance, where generated outputs may influence legal decisions, clinical care, and access to essential services (Siino et al., 2025; Costa-Gomes et al., 2026; Chen et al., 2024). Unlike open-domain conversational settings, these domains require reasoning grounded in external evidence and constrained by domain-specific rules. In such settings, useful systems must not only produce correct answers, but also remain faithful to the provided context and refuse to answer when the available evidence is insufficient (Pipitone & Alami, 2024; Peng et al., 2025). However, current language models are primarily optimised for fluent and helpful interaction, often

ICML
International Conference
On Machine Learning

EMNLP2026
Budapest, Hungary

ACCEPTED!

UNDER SUBMISSION

Building Reward Models for Grounded Legal Reasoning

Rilton Franzone [*1] Valentin Noël [*2] Puyu Wang [3 4] Philip Torr [3] Fabio J. Fehr [3 4]

Benchmark Code

Abstract

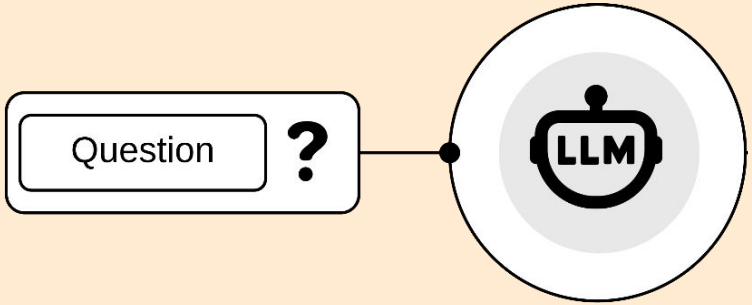
Large language models are increasingly used in high-stakes domains such as law, where systems must ground their reasoning in evidence and abstain under uncertainty. However, existing reward models prioritise helpfulness and fluency rather than contextual grounding, often producing unsupported legal reasoning in retrieval-augmented generation (RAG) settings. We introduce LEGAL-REWARDBENCH (LRB), a legal contextual reward modelling benchmark constructed from legal QA datasets for evaluating grounded legal

1. Introduction

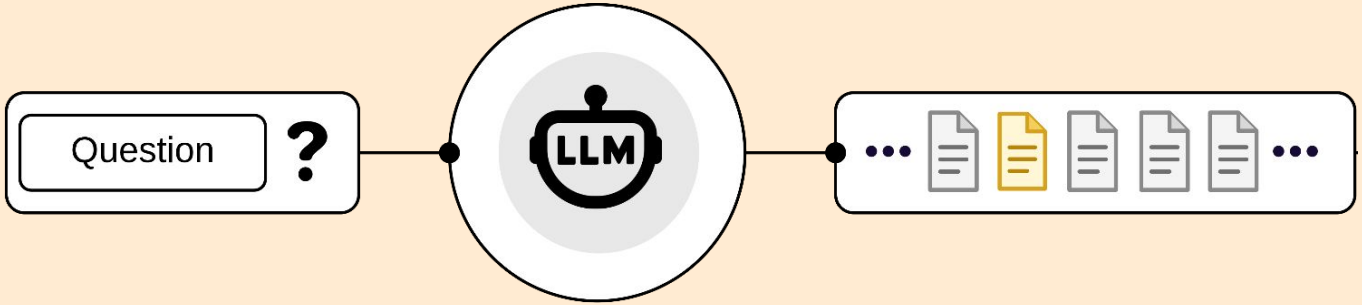
Large language models (LLMs) are increasingly deployed in high-stakes domains such as law, medicine, and finance, where generated outputs may influence legal decisions, clinical care, and access to essential services (Siino et al., 2025; Costa-Gomes et al., 2026; Chen et al., 2024). Unlike open-domain conversational settings, these domains require reasoning grounded in external evidence and constrained by domain-specific rules. In such settings, useful systems must not only produce correct answers, but also remain faithful to the provided context and refuse to answer when the available evidence is insufficient (Pipitone & Alami, 2024; Peng et al., 2025). However, current language models are primarily optimised for fluent and helpful interaction, often

Retrieval Augmented Generation

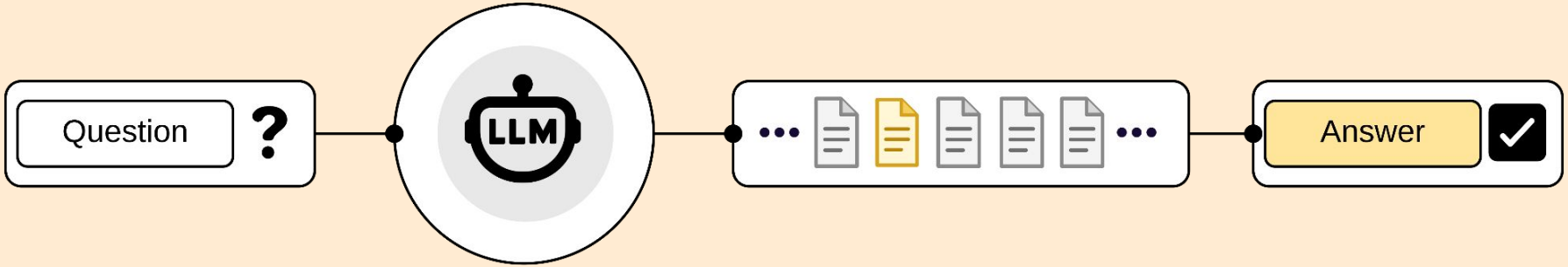
Retrieval Augmented Generation



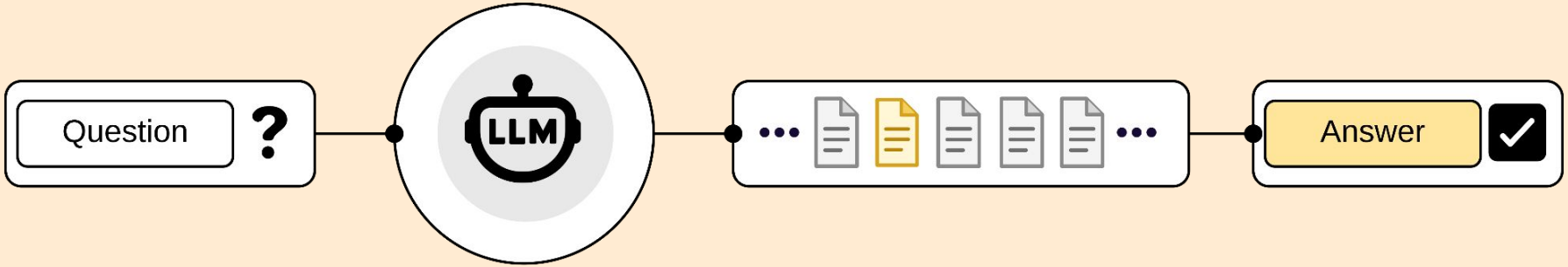
Retrieval Augmented Generation



Retrieval Augmented Generation

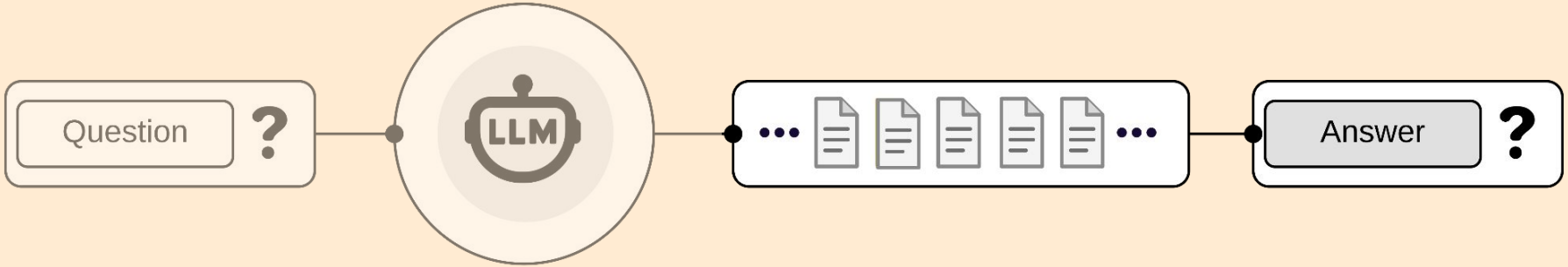


Retrieval Augmented Generation



Problems?

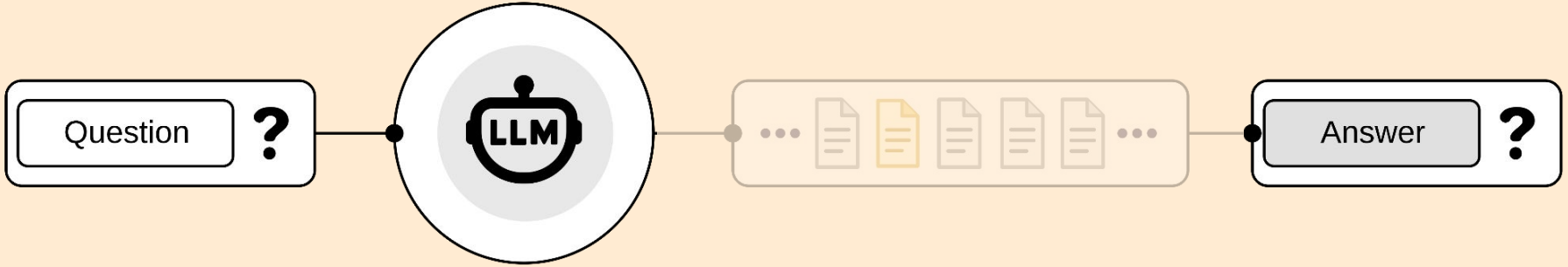
Retrieval Augmented Generation



Problems?

Retrieval: Answer or refuse?

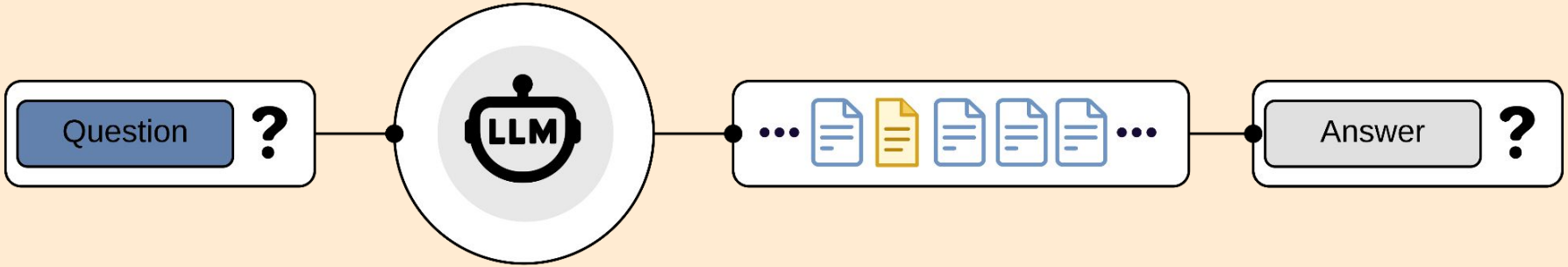
Retrieval Augmented Generation



Problems?

Hallucination: Not **grounded** in context

Retrieval Augmented Generation

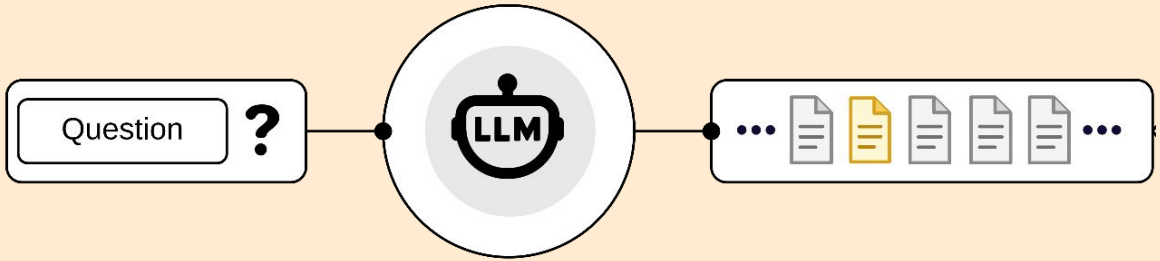


Problems?

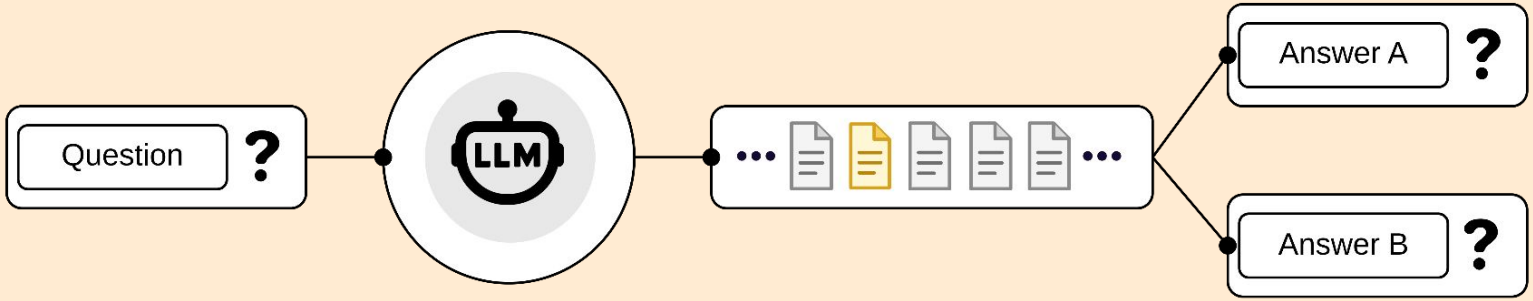
Generalisation: Different jurisdictions

We need **grounded** and **generalisable** legal reasoning

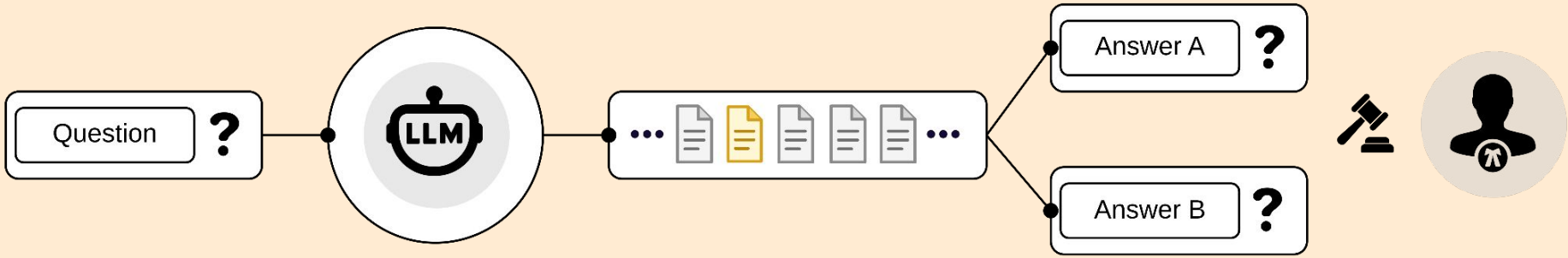
Reinforcement Learning



Reinforcement Learning

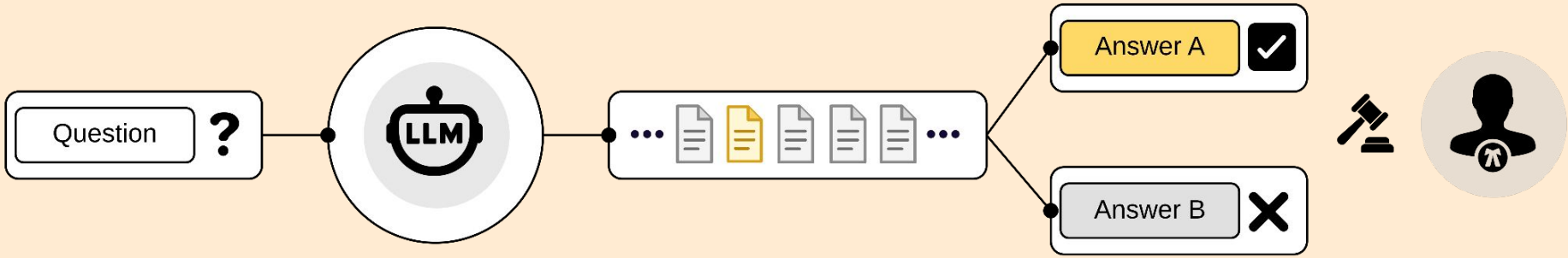


Reinforcement Learning



Reinforcement Learning

Reinforcement Learning



Problems?

RL Human Feedback: expensive + scaling issues

Reinforcement Learning

Reinforcement Learning

Reward Model

RL AI Feedback: inexpensive + scales!

How to build **reward models** for **grounded** and **generalisable** legal reasoning?

How to build **reward models** for **grounded** and **generalisable** legal reasoning?

Building Reward Models for Grounded Legal Reasoning

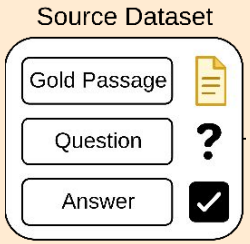
Rilton Franzone [*1] Valentin Noël [*2] Puyu Wang [3,4] Philip Torr [3] Fabio J. Fehr [3,4]

Code

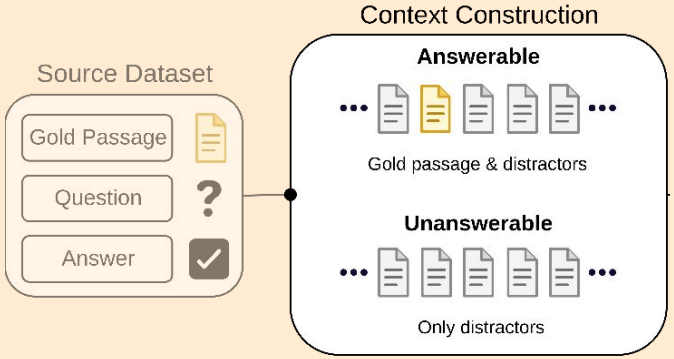
Benchmark

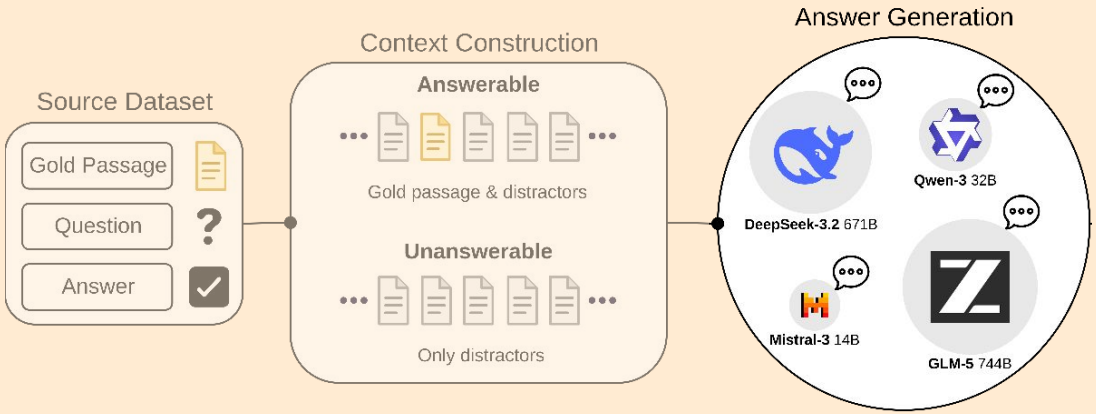
1. Introduction

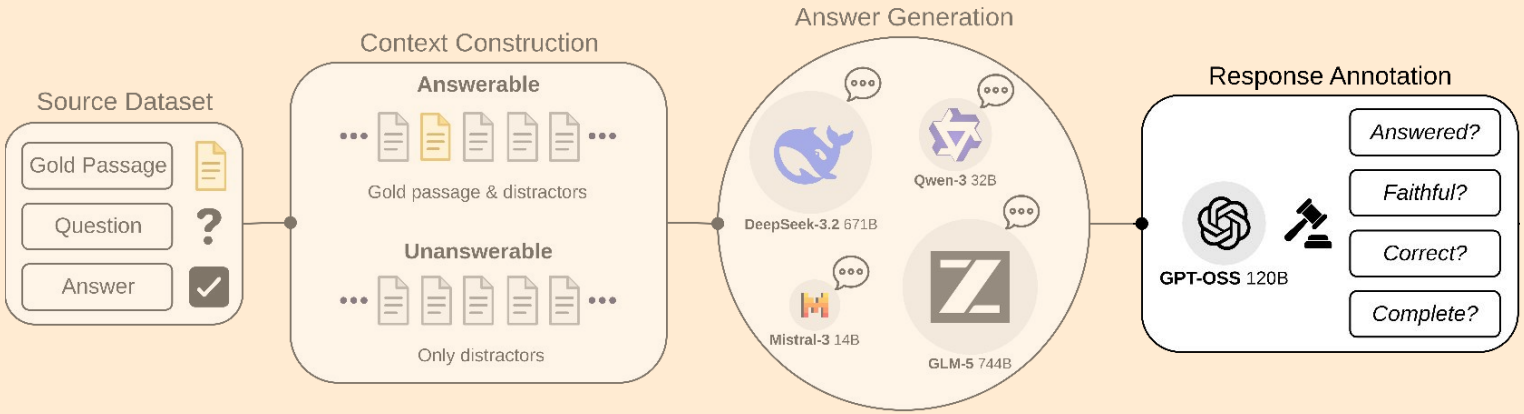
Large language models (LLMs) are increasingly deployed in high-stakes domains such as law, medicine, and finance, where generated outputs may influence legal decisions, clin-

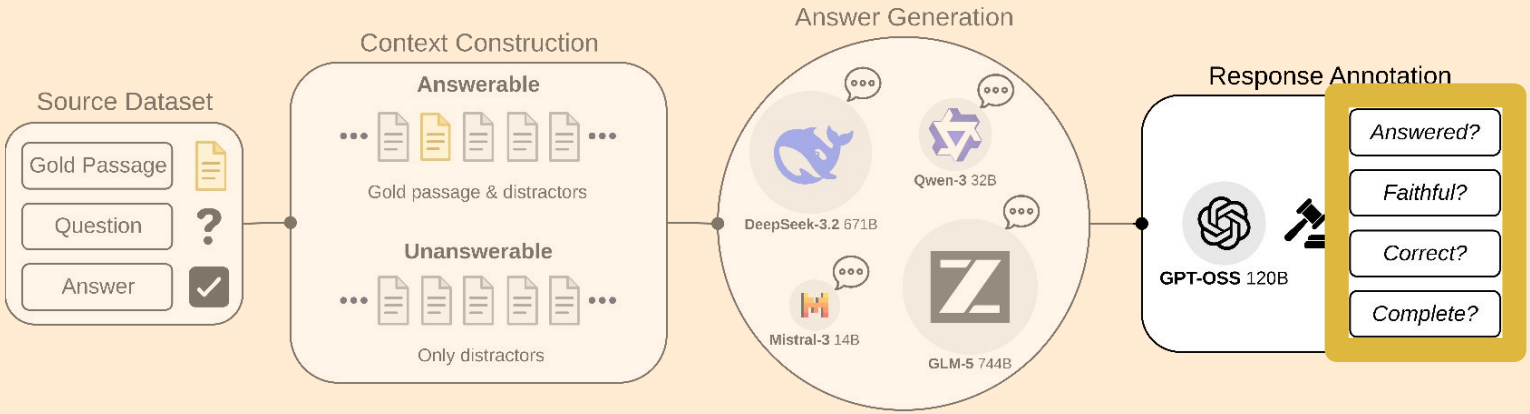


Legal RAG Bench (Butler and Butler, 2026) Victorian Criminal Law

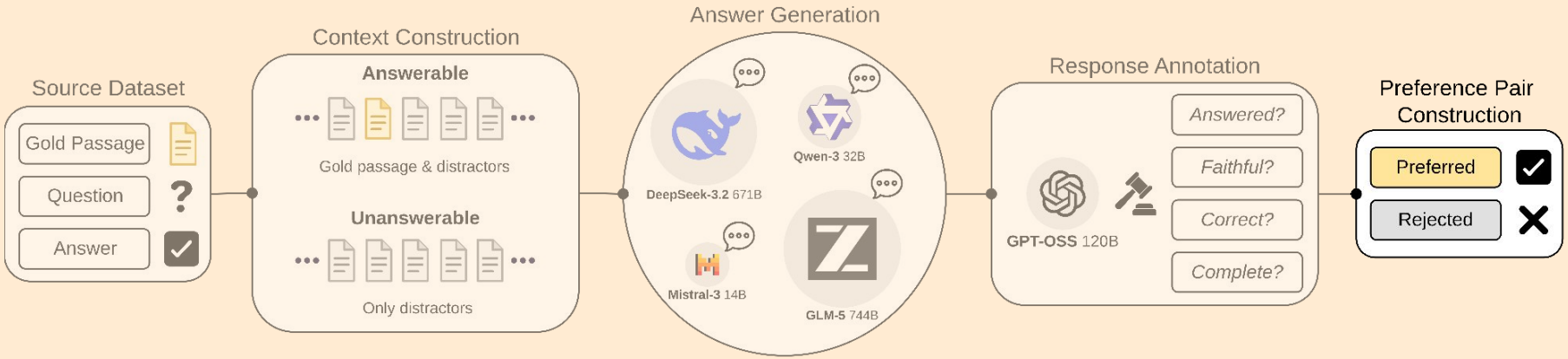

Source Dataset
Gold Passage
Question
Answer
Context Construction
Answerable
Gold passage & distractors
Unanswerable
Only distractors

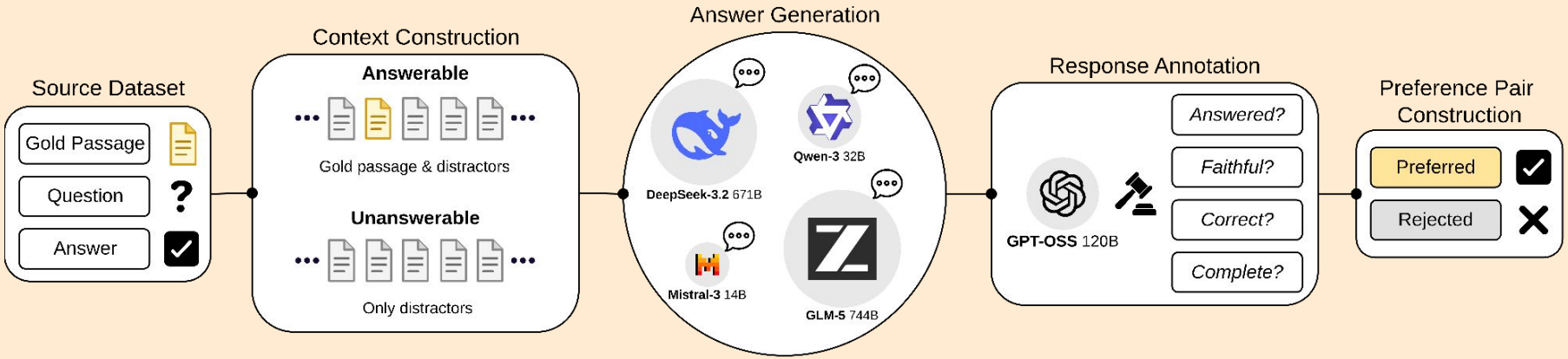

Source Dataset
Gold Passage
Question
Answer
Context Construction
Answerable
Gold passage & distractors
Unanswerable
Only distractors
Answer Generation
DeepSeek-3.2 671B
Qwen-3 32B
Mistral-3 14B
GLM-5 744B


Source Dataset
Gold Passage
Question
Answer
Context Construction
Answerable
Gold passage & distractors
Unanswerable
Only distractors
Answer Generation
DeepSeek-3.2 671B
Qwen-3 32B
Mistral-3 14B
GLM-5 744B
Response Annotation
GPT-OSS 120B
Answered?
Faithful?
Correct?
Complete?



Contextual Judge Bench (Xu et al., 2025)


Source Dataset
Gold Passage
Question
Answer
Context Construction
Answerable
Gold passage & distractors
Unanswerable
Only distractors
Answer Generation
DeepSeek-3.2 671B
Qwen-3 32B
Mistral-3 14B
GLM-5 744B
Response Annotation
GPT-OSS 120B
Answered?
Faithful?
Correct?
Complete?
Preference Pair Construction
Preferred
Rejected


Source Dataset
Gold Passage
Question
Answer
Context Construction
Answerable
Gold passage & distractors
Unanswerable
Only distractors
Answer Generation
DeepSeek-3.2 671B
Qwen-3 32B
Mistral-3 14B
GLM-5 744B
Response Annotation
GPT-OSS 120B
Answered?
Faithful?
Correct?
Complete?
Preference Pair Construction
Preferred
Rejected

Legal Reward Bench

Failure Type
Refusal
Faithfulness
Correctness
Completeness

Legal Reward Bench

Failure Type	Question q
Refusal	What sentence was imposed on the accused?
Faithfulness	
Correctness	
Completeness	

Legal Reward Bench

Failure Type	Question q	Preferred Response r^+
Refusal	What sentence was imposed on the accused?	"The provided context is insufficient to answer this question."
Faithfulness		
Correctness		
Completeness		

Legal Reward Bench

Failure Type	Question q	Preferred Response r^+	Rejected Response r^-
Refusal	What sentence was imposed on the accused?	"The provided context is insufficient to answer this question."	"The accused was sentenced to five years imprisonment."
Faithfulness			
Correctness			
Completeness			

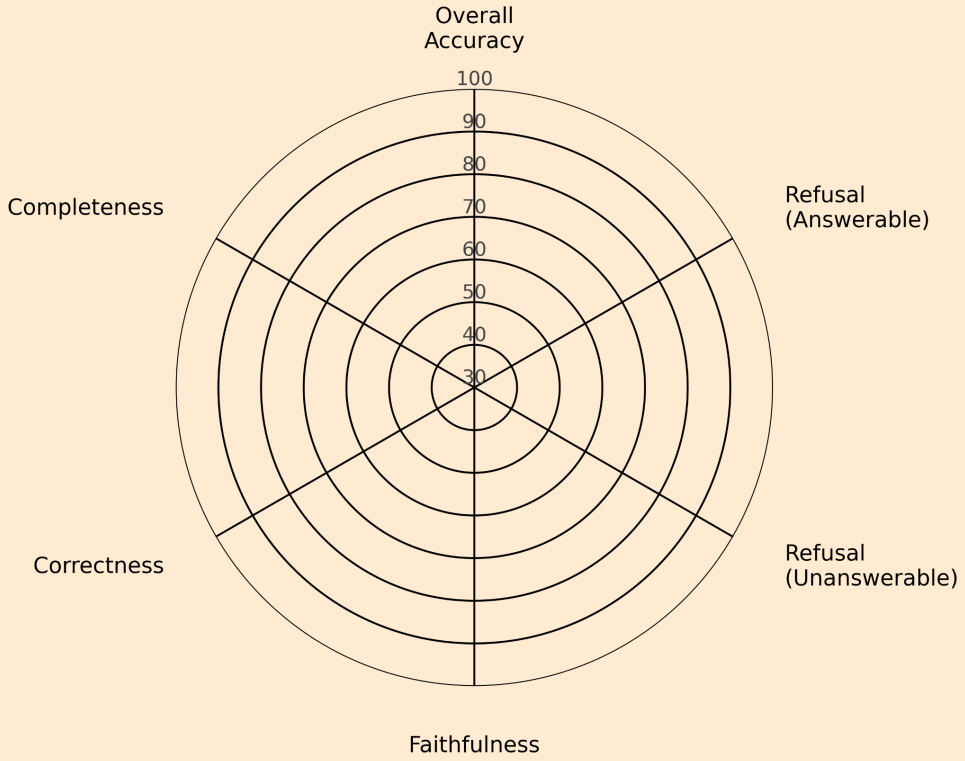
Legal Reward Bench

Failure Type	Question q	Preferred Response r^+	Rejected Response r^-
Refusal	What sentence was imposed on the accused?	"The provided context is insufficient to answer this question."	"The accused was sentenced to five years imprisonment."
Faithfulness	Which Act introduced cyberstalking offences?	"The Crimes (Stalking) Act 2003 introduced cyberstalking offences."	"The Crimes Amendment Act 2001 introduced cyberstalking offences."
Correctness	Does Victoria protect against double jeopardy?	"Yes. Victoria protects against double jeopardy."	"No. Victoria does not recognise double jeopardy protections."
Completeness	Must jurors be excused after seeing pre-trial publicity?	"No. Judges can usually address prejudice through directions to the jury."	"No."

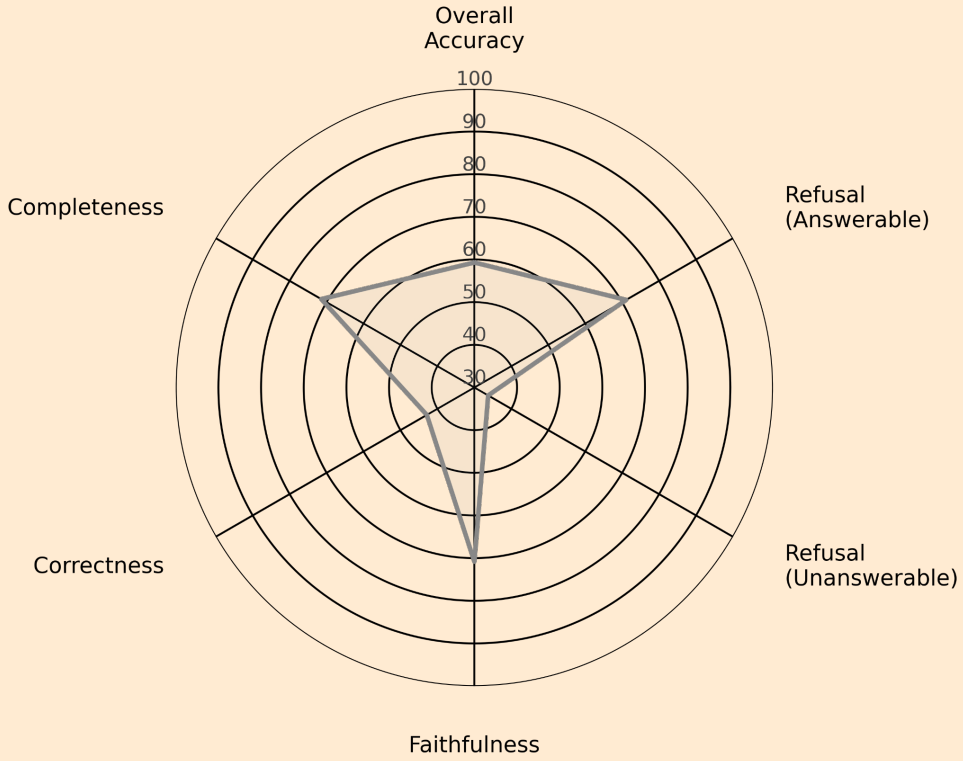
Legal Reward Bench

Failure Type	Question q	Preferred Response r^+	Rejected Response r^-
Refusal	What sentence was imposed on the accused?	"The provided context is insufficient to answer this question."	"The accused was sentenced to five years imprisonment."
Faithfulness	Which Act introduced cyberstalking offences?	"The Crimes (Stalking) Act 2003 introduced cyberstalking offences."	"The Crimes Amendment Act 2001 introduced cyberstalking offences."
Correctness	Does Victoria protect against double jeopardy?	"Yes. Victoria protects against double jeopardy."	"No. Victoria does not recognise double jeopardy protections."
Completeness	Must jurors be excused after seeing pre-trial publicity?	"No. Judges can usually address prejudice through directions to the jury."	"No."

Fixed words + Shorter lengths


Overall
Accuracy
100
90
80
70
60
50
40
30
Refusal
(Answerable)
Refusal
(Unanswerable)
Faithfulness
Correctness
Completeness

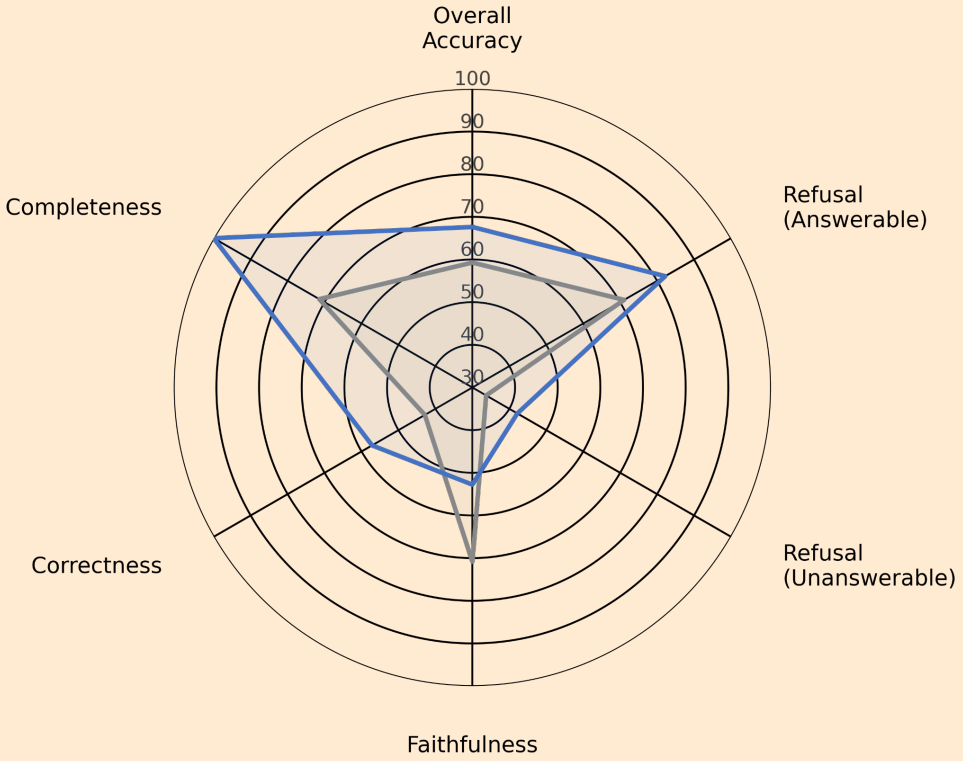
Ministral-8B



Ministral-8B

+ DPO CBJ

Contextual Judge Bench (Xu et al., 2025)



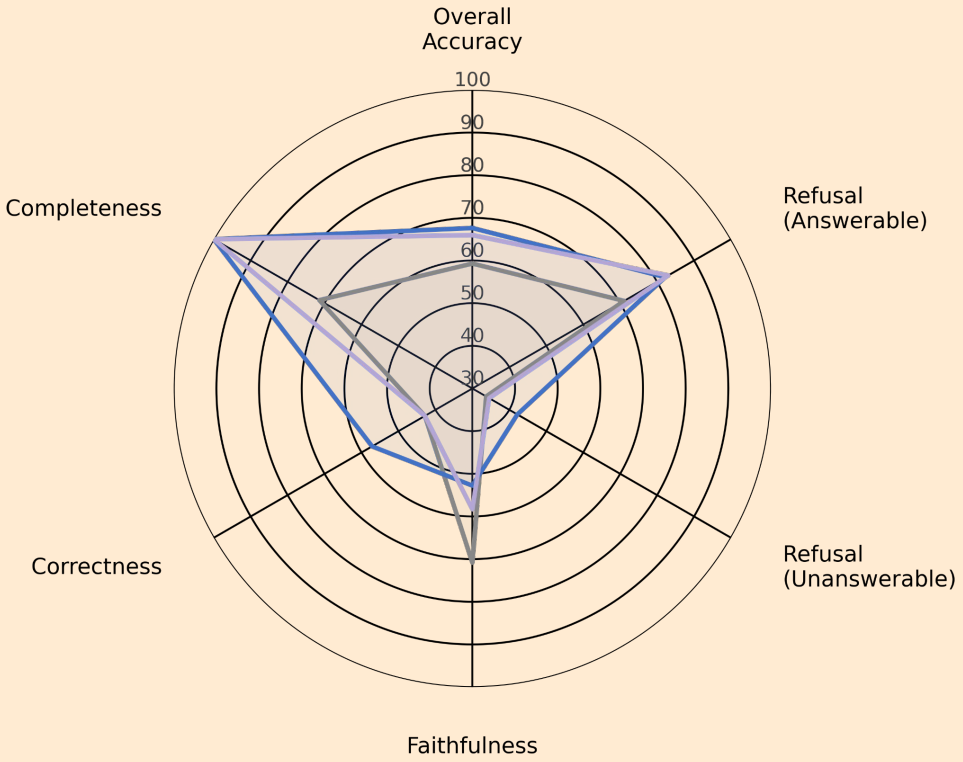
Ministral-8B

+ **DPO CBJ**

  Contextual Judge Bench (Xu et al., 2025)

+ **DPO CBJ + LRB**

  Contextual Judge Bench (Xu et al., 2025)
  Legal Reward Bench (Ours)



Legal Reward Bench

Failure Type	Question q	Preferred Response r^+	Rejected Response r^-
Refusal	What sentence was imposed on the accused?	"The provided context is insufficient to answer this question."	"The accused was sentenced to five years imprisonment."
Faithfulness	Which Act introduced cyberstalking offences?	"The Crimes (Stalking) Act 2003 introduced cyberstalking offences."	"The Crimes Amendment Act 2001 introduced cyberstalking offences."
Correctness	Does Victoria protect against double jeopardy?	"Yes. Victoria protects against double jeopardy."	"No. Victoria does not recognise double jeopardy protections."
Completeness	Must jurors be excused after seeing pre-trial publicity?	"No. Judges can usually address prejudice through directions to the jury."	"No."

Reward-Hacking

Legal Reward Bench

Failure Type	Question q	Preferred Response r^+	Rejected Response r^-
Refusal	What sentence was imposed on the accused?	"The provided context is insufficient to answer this question."	"The accused was sentenced to five years imprisonment."
Faithfulness	Which Act introduced cyberstalking offences?	"The Crimes (Stalking) Act 2003 introduced cyberstalking offences."	"The Crimes Amendment Act 2001 introduced cyberstalking offences."
Correctness	Does Victoria protect against double jeopardy?	"Yes. Victoria protects against double jeopardy."	"No. Victoria does not recognise double jeopardy protections."
Completeness	Must jurors be excused after seeing pre-trial publicity?	"No. Judges can usually address prejudice through directions to the jury."	"No." Judges must follow regulations.

Reward-Hacking

Length balancing augmentation

Ministral-8B

+ **DPO CBJ**

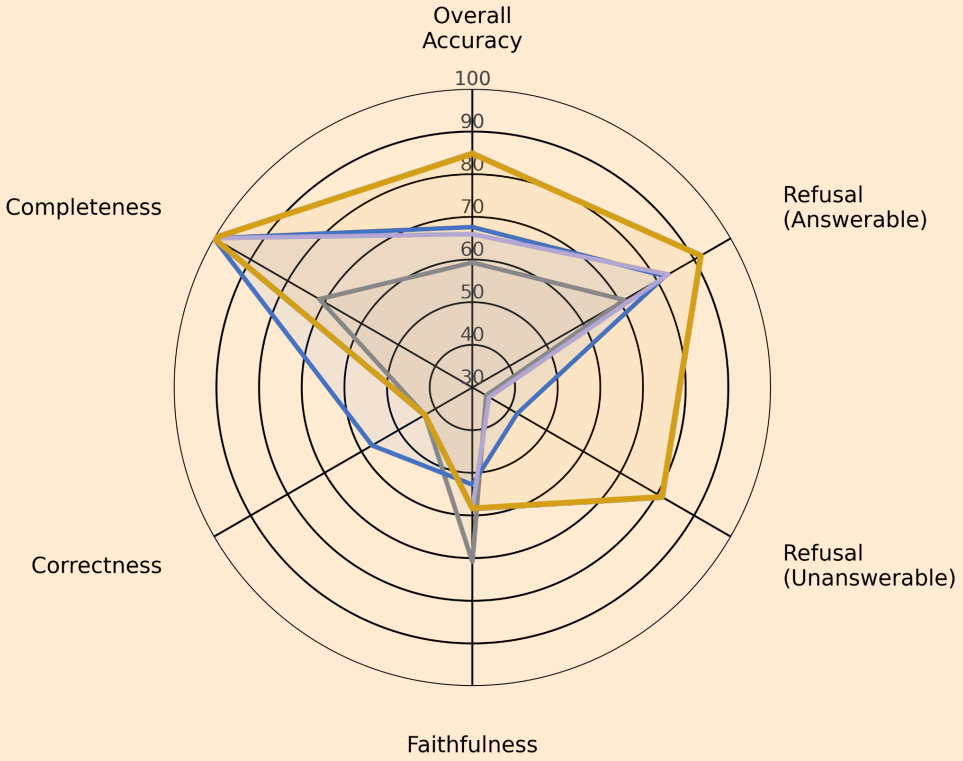
  Contextual Judge Bench (Xu et al., 2025)

+ **DPO CBJ + LRB**

  Contextual Judge Bench (Xu et al., 2025)
  Legal Reward Bench (Ours)

+ **DPO CBJ + LRB Length Aug**

  Contextual Judge Bench (Xu et al., 2025)
  Legal Reward Bench (Ours)
  Length Augmentation



Ministral-8B

+ **DPO CBJ**

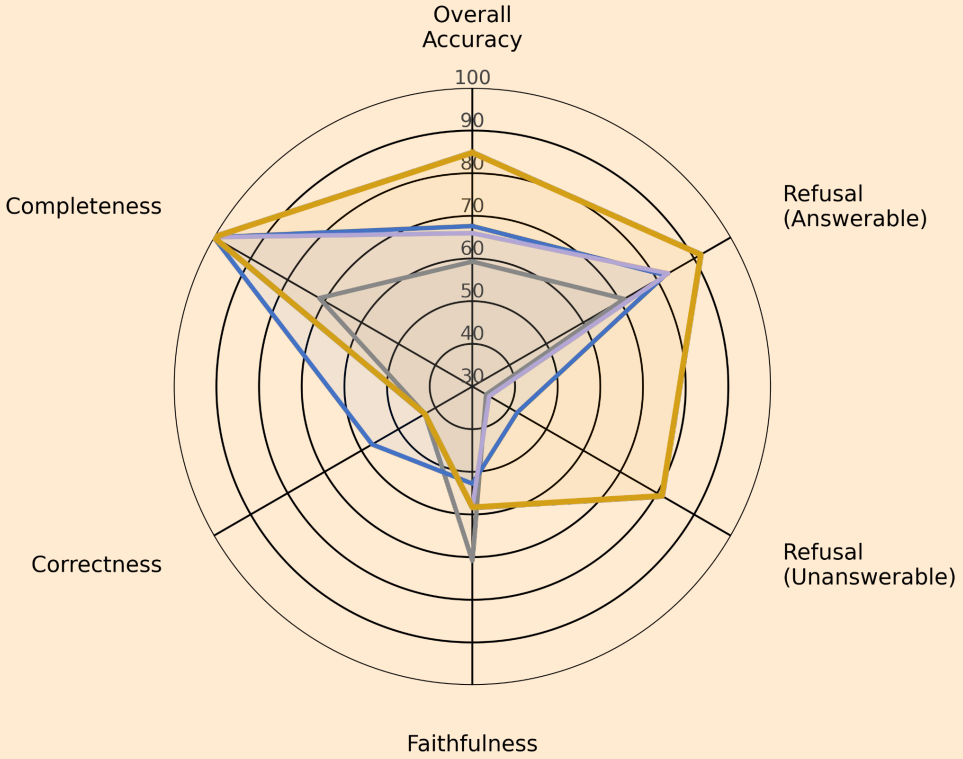
  Contextual Judge Bench (Xu et al., 2025)

+ **DPO CBJ + LRB**

  Contextual Judge Bench (Xu et al., 2025)
  Legal Reward Bench (Ours)

+ **DPO CBJ + LRB Length Aug**

  Contextual Judge Bench (Xu et al., 2025)
  Legal Reward Bench (Ours)
  Length Augmentation



Sensitive to Reward-Hacking

Ministral-8B

+ **DPO CBJ**

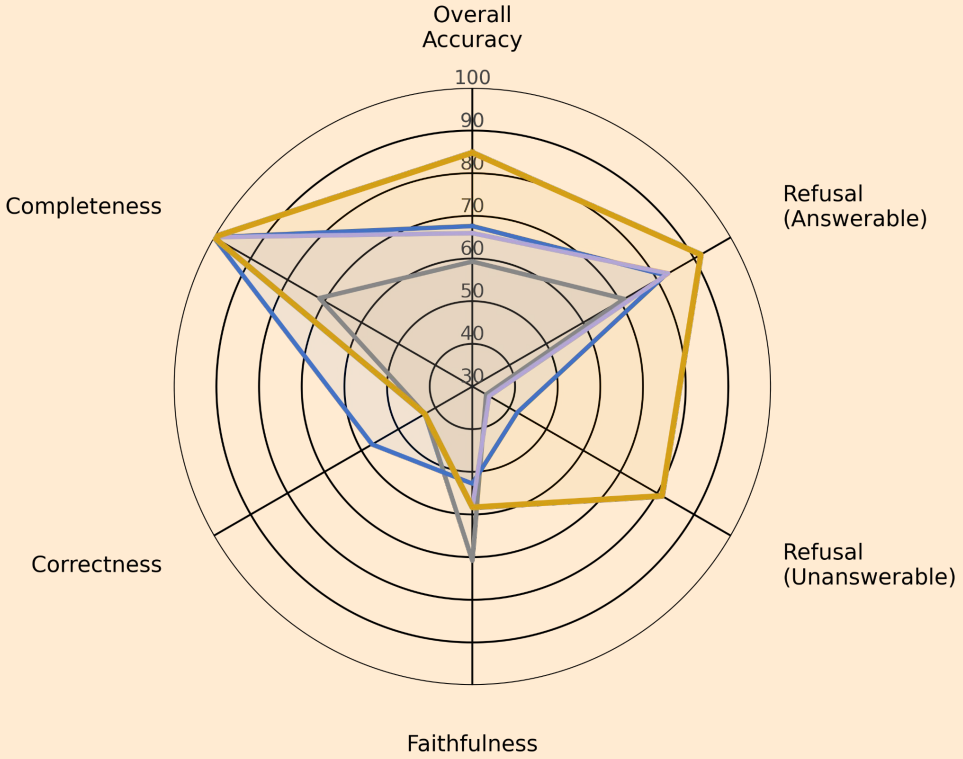
Contextual Judge Bench (Xu et al., 2025)

+ **DPO CBJ + LRB**

Contextual Judge Bench (Xu et al., 2025)
Legal Reward Bench (Ours)

+ **DPO CBJ + LRB Length Aug**

Contextual Judge Bench (Xu et al., 2025)
Legal Reward Bench (Ours)
Length Augmentation



Grounded Legal Reward Models

Bar Exam: US Bar exam MCQ (Zheng et al.,2025)

Housing: US state housing statutes MCQ (Zheng et al.,2025)

Bar Exam: US Bar exam MCQ (Zheng et al.,2025)

Housing: US state housing statutes MCQ (Zheng et al.,2025)

Model	Bar Exam	Δ	Housing	Δ
Random	50.0	–	50.0	–
Ministral-8B	56.4 [47,65]	–	54.2 [50,58]	–

Bar Exam: US Bar exam MCQ (Zheng et al.,2025)

Housing: US state housing statutes MCQ (Zheng et al.,2025)

Model	Bar Exam	Δ	Housing	Δ
Random	50.0	–	50.0	–
Ministral-8B	56.4 [47,65]	–	54.2 [50,58]	–
+DPO	60.7 [52,69]	+4.3	**70.4** [66,74]	**+16.2**

Bar Exam: US Bar exam MCQ (Zheng et al.,2025)

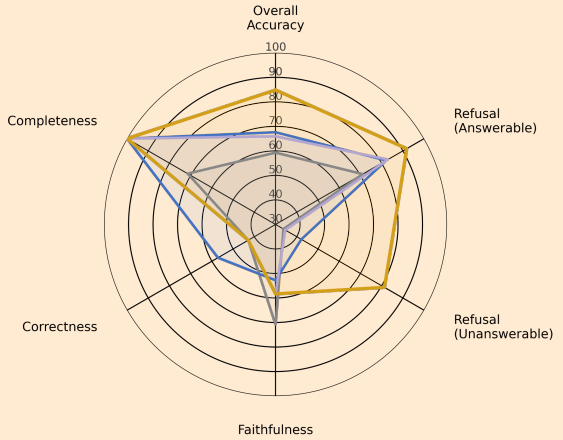
Housing: US state housing statutes MCQ (Zheng et al.,2025)

Model	Bar Exam	Δ	Housing	Δ
Random	50.0	–	50.0	–
Ministral-8B	56.4 [47,65]	–	54.2 [50,58]	–
+DPO	60.7 [52,69]	+4.3	**70.4** [66,74]	**+16.2**

Cross-Jurisdiction **Generalisation**

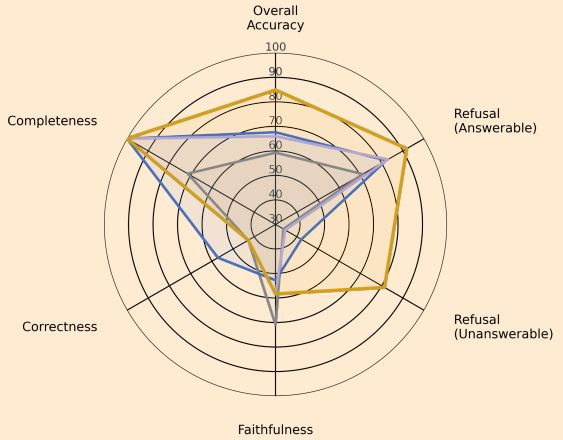
How to build **reward models** for **grounded** and **generalisable** legal reasoning

Grounded Legal Reward Models



Grounded Legal Reward Models

Sensitive to **Reward-Hacking**

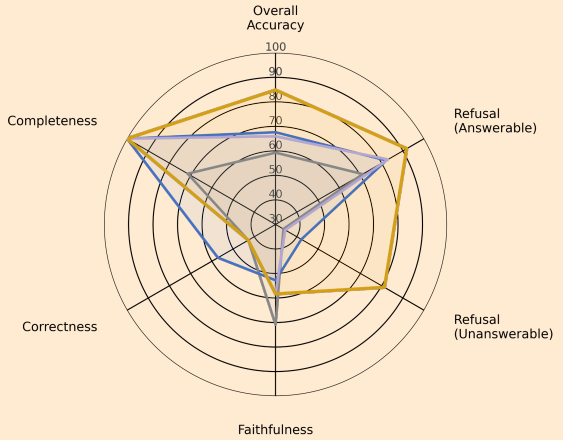


Failure Type	Question q	Preferred Response r^+	Rejected Response r^-
Refusal	What sentence was imposed on the accused?	"The provided context is insufficient to answer this question."	"The accused was sentenced to five years imprisonment."
Faithfulness	Which Act introduced cyberstalking offences?	"The Crimes (Stalking) Act 2003 introduced cyberstalking offences."	"The Crimes Amendment Act 2001 introduced cyberstalking offences."
Correctness	Does Victoria protect against double jeopardy?	"Yes. Victoria protects against double jeopardy."	"No. Victoria does not recognise double jeopardy protections."
Completeness	Must jurors be excused after seeing pre-trial publicity?	"No. Judges can usually address prejudice through directions to the jury."	"No."

Grounded Legal Reward Models

Sensitive to **Reward-Hacking**

Cross-Jurisdiction **Generalisation**



Failure Type	Question q	Preferred Response r^+	Rejected Response r^-
Refusal	What sentence was imposed on the accused?	"The provided context is insufficient to answer this question."	"The accused was sentenced to five years imprisonment."
Faithfulness	Which Act introduced cyberstalking offences?	"The Crimes (Stalking) Act 2003 introduced cyberstalking offences."	"The Crimes Amendment Act 2001 introduced cyberstalking offences."
Correctness	Does Victoria protect against double jeopardy?	"Yes. Victoria protects against double jeopardy."	"No. Victoria does not recognise double jeopardy protections."
Completeness	Must jurors be excused after seeing pre-trial publicity?	"No. Judges can usually address prejudice through directions to the jury."	"No."

Model	Bar Exam	Δ	Housing	Δ
Random	50.0	–	50.0	–
Ministral-8B	56.4 [47,65]	–	54.2 [50,58]	–
+DPO	60.7 [52,69]	+4.3	**70.4** [66,74]	**+16.2**