Measuring and Enhancing Trustworthiness of LLMs in RAG

DeCLaRe Lab

What is a good quality output for RAG?
How does one measure a “good” output for RAG?

What is a good quality output for RAG?
How does one measure a “good” output for RAG?

Loosely, it should be accurate and “grounded” in the documents

What is a good quality output for RAG?

Where is President Barack Obama born?

[1] Barack Obama was born on August 4, 1961...

[2] He was born to an 18-year-old American mother ...

[3] At six, Obama and his mother moved to Indonesia to join his stepfather.

What is a good quality output for RAG?

Where is President Barack Obama born?

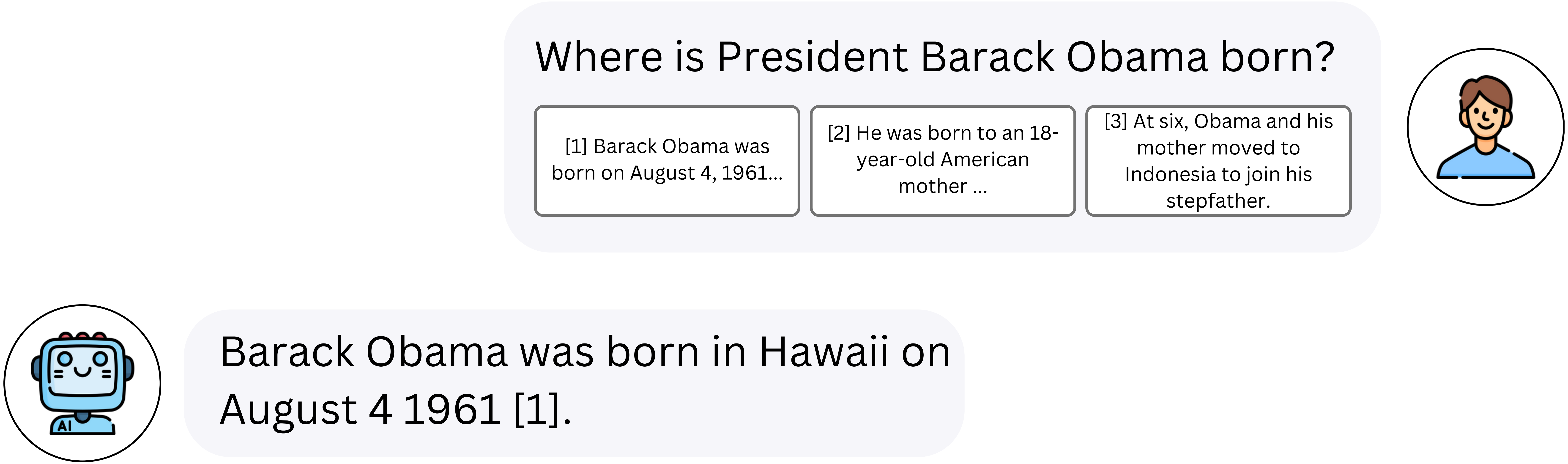
[1] Barack Obama was born on August 4, 1961...

[2] He was born to an 18-year-old American mother ...

[3] At six, Obama and his mother moved to Indonesia to join his stepfather.

Barack Obama was born in Hawaii on August 4 1961 [1].

What is a good quality output for RAG?



Where is President Barack Obama born?

[1] Barack Obama was born on August 4, 1961...

[2] He was born to an 18-year-old American mother ...

[3] At six, Obama and his mother moved to Indonesia to join his stepfather.

Barack Obama was born in Hawaii on August 4 1961 [1].

Is this a good output?

What is a good quality output for RAG?

Where is President Barack Obama born?

[1] Barack Obama was born on August 4, 1961...

[2] He was born to an 18-year-old American mother ...

[3] At six, Obama and his mother moved to Indonesia to join his stepfather.

Barack Obama was born in Hawaii on August 4 1961 [1].

- Doc 1 supports the date of birth
- But the location of birth, the main answer to the question, is unsupported

What is a good quality output for RAG?

Where is President Barack Obama born?

[1] Barack Obama was born on August 4, 1961...

[2] He was born to an 18-year-old American mother ...

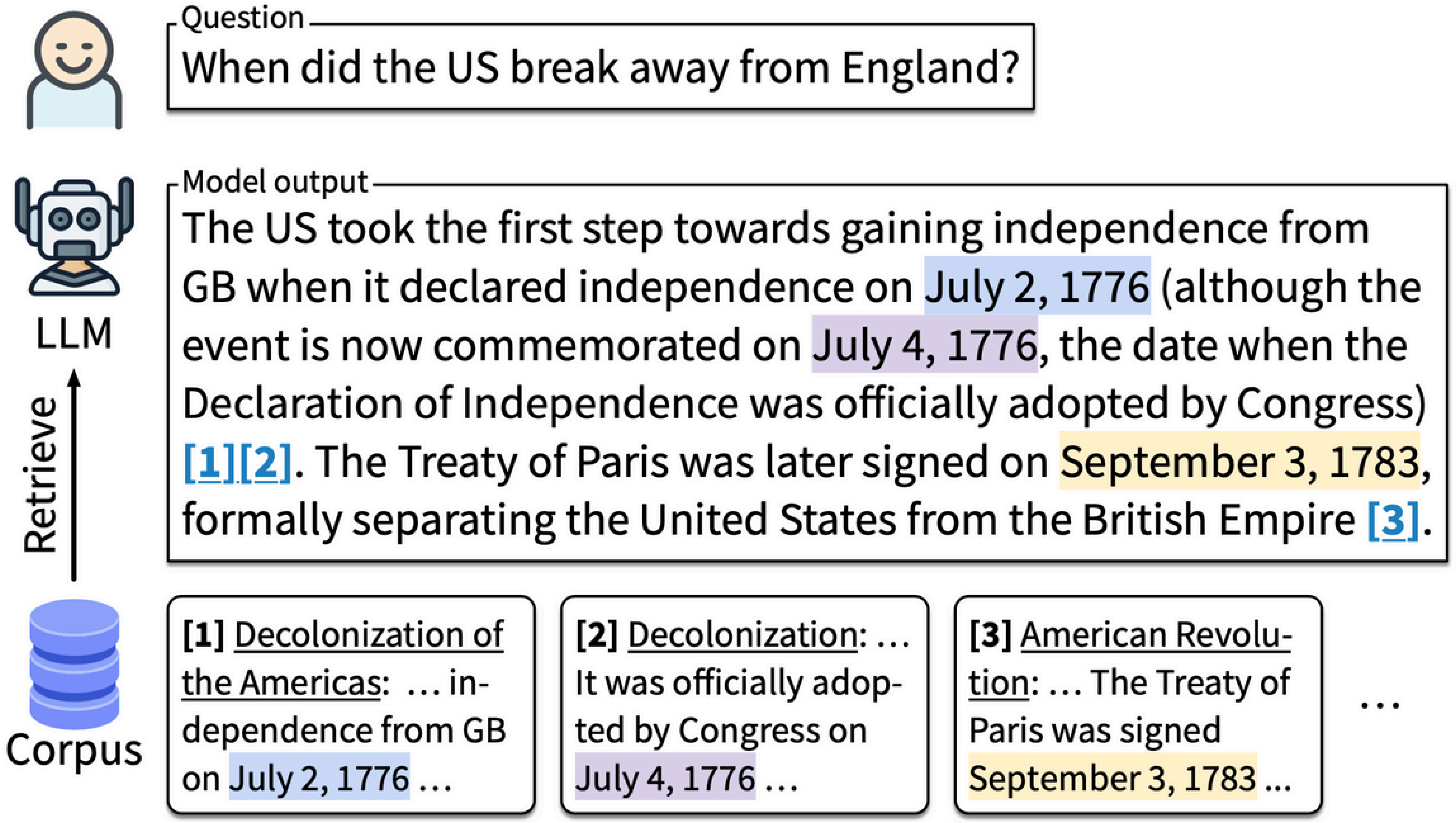
[3] At six, Obama and his mother moved to Indonesia to join his stepfather.

None of the documents contain information on **where** Obama was born.

Barack Obama was born in Hawaii on August 4 1961 [1].

LLM should have **refused to answer** due to **insufficient information**

Previous Works



When did the US break away from England?

The US took the first step towards gaining independence from GB when it declared independence on July 2, 1776 (although the event is now commemorated on July 4, 1776, the date when the Declaration of Independence was officially adopted by Congress) [1][2]. The Treaty of Paris was later signed on September 3, 1783, formally separating the United States from the British Empire [3].

[1] Decolonization of the Americas: … independence from GB on July 2, 1776 …

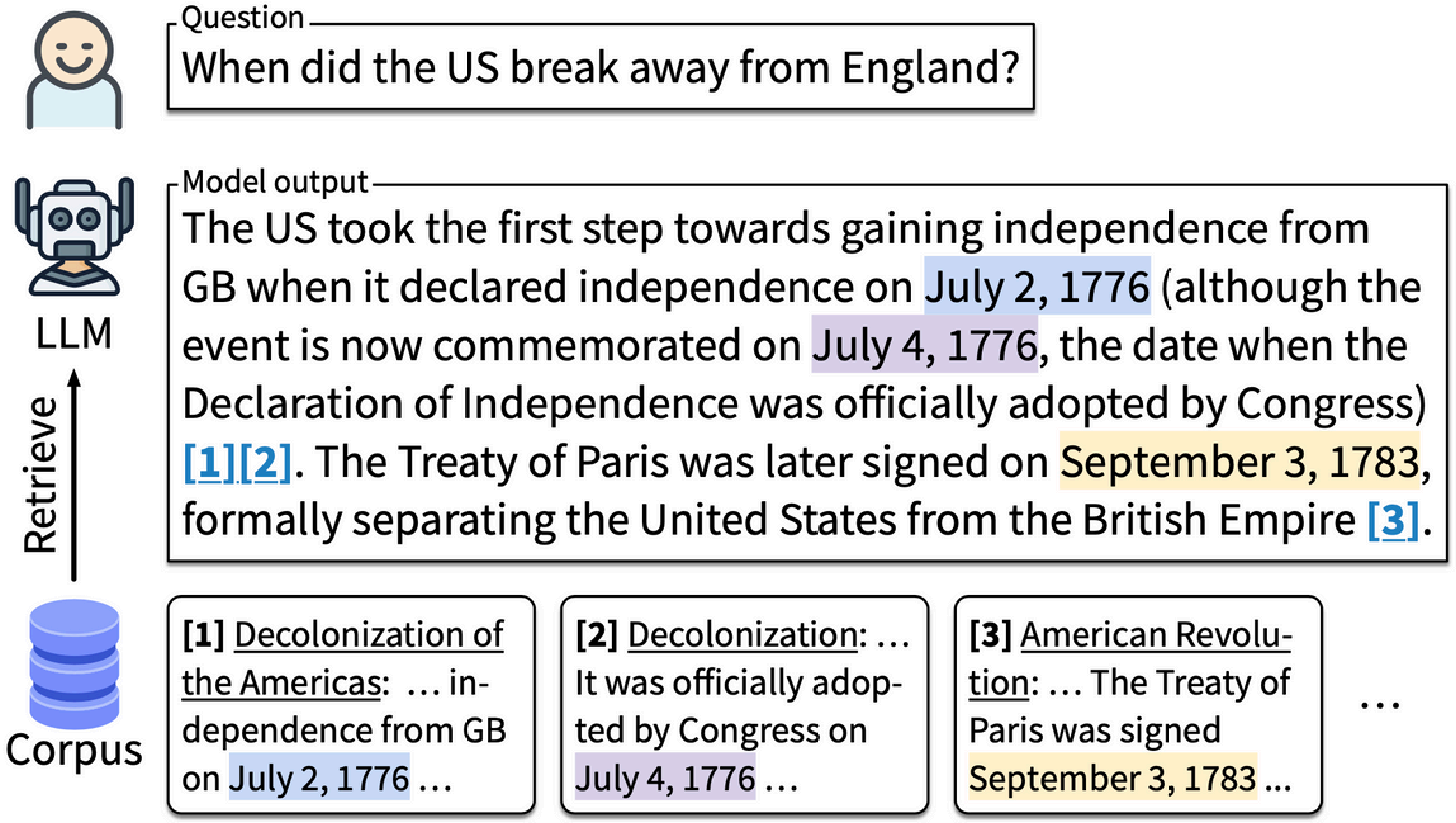
[2] Decolonization: … It was officially adopted by Congress on July 4, 1776 …

[3] American Revolution: … The Treaty of Paris was signed September 3, 1783 ...

https://github.com/princeton-nlp/ALCE

- **Ideal**: Make LLM ground answer in external documents, rather than parametric knowledge
- **Incumbent Definition of Grounded**: citations support claims sufficiently and precisely => Under such a definition of groundedness, the previous response would have been deemed as good!

Previous Works



When did the US break away from England?

The US took the first step towards gaining independence from GB when it declared independence on July 2, 1776 (although the event is now commemorated on July 4, 1776, the date when the Declaration of Independence was officially adopted by Congress) [1][2]. The Treaty of Paris was later signed on September 3, 1783, formally separating the United States from the British Empire [3].

[1] Decolonization of the Americas: … independence from GB on July 2, 1776 …

[2] Decolonization: … It was officially adopted by Congress on July 4, 1776 …

[3] American Revolution: … The Treaty of Paris was signed September 3, 1783 ...

https://github.com/princeton-nlp/ALCE

Furthemore...

- Under such a case, bad retriever = bad outputs = bad scores. Search system is measured as a whole
- Measure model's effectiveness in the RAG system without the confounding effect of the retriever
- Decouple the influence of model behaviour and search efficacy on overall system performance

Key Contributions

1 Introduced and measured a more robust and holistic definition of groundedness

2 A metric, **Trust-Score**, to specifically measure LLM groundedness in RAG systems

3 An alignment approach, **Trust-Align** to enforce LLM groundedness

LLM Groundedness

Grounded response:

✅ **Refuse to answer** questions whose answer that cannot be verified

✅ **Correctly answers** question using in-context documents

✅ **Inline citations** to the in-context documents to support generated answers

Trust-Score

Response

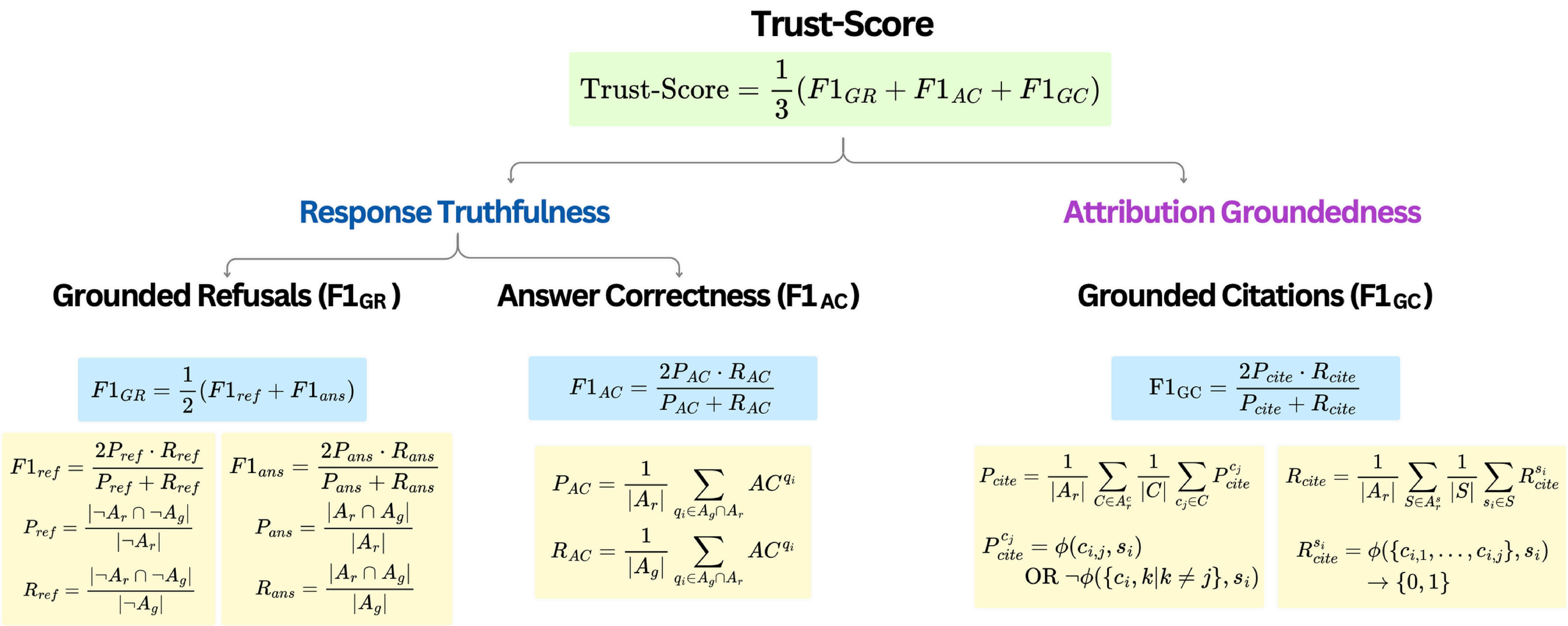
1 **Grounded Refusals**: Is the model able to discern which questions can be answered or refused based on the provided documents?

2 **Exact Match scores**: For the answerable questions, is the response correct?

Citations

1 **Citation recall**: Are generated statements well-supported by the set citations?

2 **Citation precision**: Are the citations relevant to the statements?



Trust-Score

$$\text{Trust-Score} = \frac{1}{3}(F1_{GR} + F1_{AC} + F1_{GC})$$

Response Truthfulness

Grounded Refusals ($F1_{GR}$)

$$F1_{GR} = \frac{1}{2}(F1_{ref} + F1_{ans})$$

$$F1_{ref} = \frac{2P_{ref} \cdot R_{ref}}{P_{ref} + R_{ref}}$$

$$P_{ref} = \frac{|\neg A_r \cap \neg A_g|}{|\neg A_r|}$$

$$R_{ref} = \frac{|\neg A_r \cap \neg A_g|}{|\neg A_g|}$$

$$F1_{ans} = \frac{2P_{ans} \cdot R_{ans}}{P_{ans} + R_{ans}}$$

$$P_{ans} = \frac{|A_r \cap A_g|}{|A_r|}$$

$$R_{ans} = \frac{|A_r \cap A_g|}{|A_g|}$$

Answer Correctness ($F1_{AC}$)

$$F1_{AC} = \frac{2P_{AC} \cdot R_{AC}}{P_{AC} + R_{AC}}$$

$$P_{AC} = \frac{1}{|A_r|} \sum_{q_i \in A_g \cap A_r} AC^{q_i}$$

$$R_{AC} = \frac{1}{|A_g|} \sum_{q_i \in A_g \cap A_r} AC^{q_i}$$

Attribution Groundedness

Grounded Citations ($F1_{GC}$)

$$\text{F1}_{\text{GC}} = \frac{2P_{cite} \cdot R_{cite}}{P_{cite} + R_{cite}}$$

$$P_{cite} = \frac{1}{|A_r|} \sum_{C \in A_r^c} \frac{1}{|C|} \sum_{c_j \in C} P_{cite}^{c_j}$$

$$P_{cite}^{c_j} = \phi(c_{i,j}, s_i) \\ \text{OR} \ \neg\phi(\{c_i, k | k \neq j\}, s_i)$$

$$R_{cite} = \frac{1}{|A_r|} \sum_{S \in A_r^s} \frac{1}{|S|} \sum_{s_i \in S} R_{cite}^{s_i}$$

$$R_{cite}^{s_i} = \phi(\{c_{i,1}, \ldots, c_{i,j}\}, s_i) \\ \rightarrow \{0, 1\}$$

A_r Set of answered questions

A_g Set of answerable questions

A_r^c / A_r^s Set of answered questions (statements only, no citations / citations only, no statements)

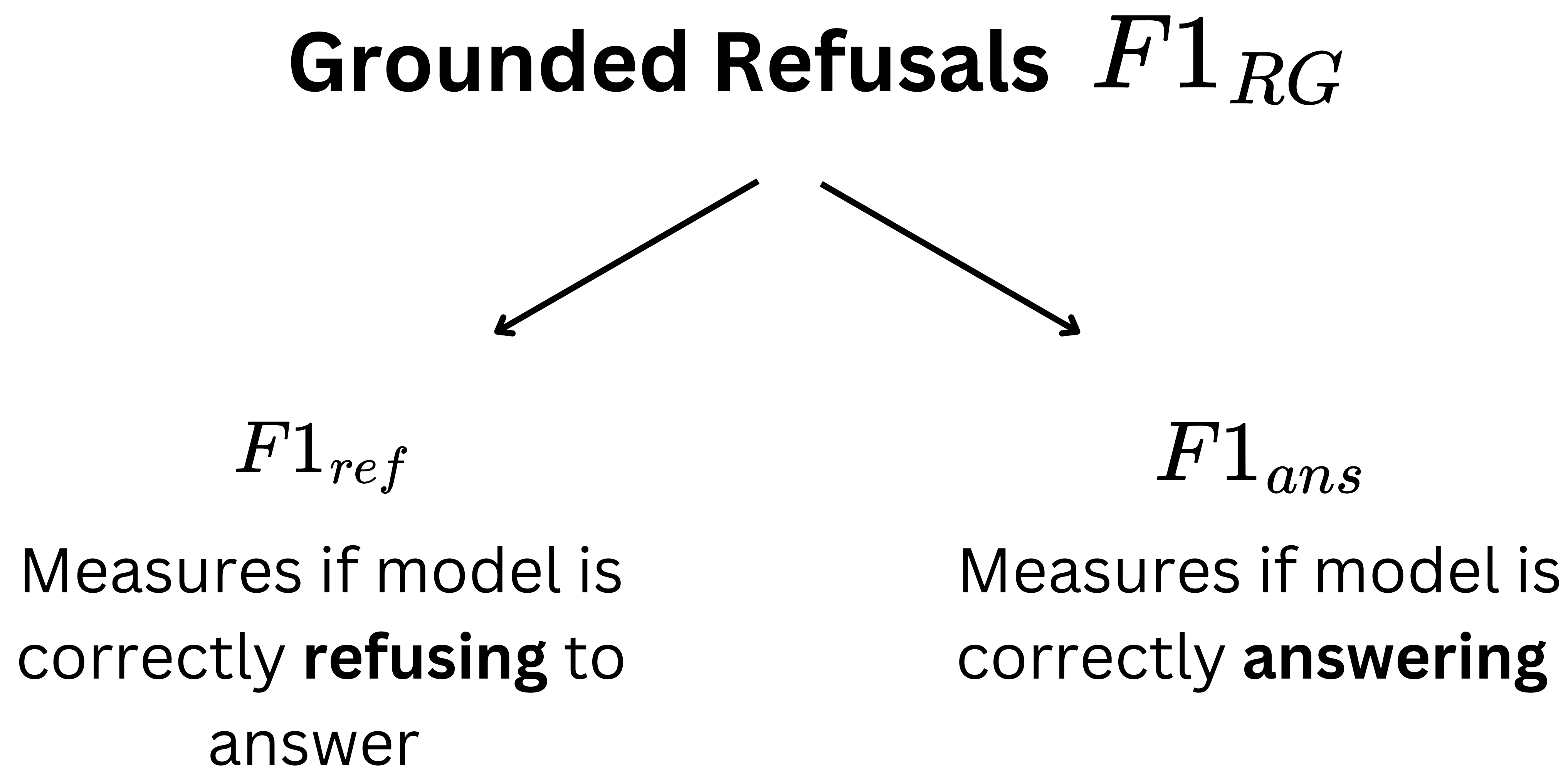
$|A|$ Number of elements in the set

$s_i, \{c_{i,1}, c_{i,2}, \ldots\}$ Statement and corresponding citations

AC^{q_i} Answer correctness for question q_i

$R_{cite}^{s_i}$ Citation recall for statement s_i

$P_{cite}^{c_j}$ Citation precision for citation c_j



Grounded Refusals $F1_{RG}$

$F1_{ref}$

Measures if model is correctly **refusing** to answer

$F1_{ans}$

Measures if model is correctly **answering**

Quality of Refusals $F1_{ref}$

$$R_{ref} = \frac{|\neg A_r \cap \neg A_g|}{|\neg A_g|}$$

What is the proportion of unanswerable questions that the model is refusing?

$$P_{ref} = \frac{|\neg A_r \cap \neg A_g|}{|\neg A_r|}$$

What is the proportion of refused questions are unanswerable?

$\neg A_r$ Set of questions where model refused to answer

$\neg A_g$ Set of questions where it is ground truth unanswerable

$$F1_{ref} = \frac{2P_{ref} \cdot R_{ref}}{P_{ref} + R_{ref}}$$

Quality of answering $F1_{ans}$

$$R_{ans} = \frac{|A_r \cap A_g|}{|A_g|}$$

What is the proportion of answerable questions that the model is answering?

$$P_{ans} = \frac{|A_r \cap A_g|}{|A_r|}$$

What is the proportion of answered questions are answerable?

A_r Set of questions where model provided an answer

A_g Set of questions where it is ground truth answerable

$$F1_{ans} = \frac{2P_{ans} \cdot R_{ans}}{P_{ans} + R_{ans}}$$

Grounded Refusals $F1_{RG}$

$$F1_{RG} = \frac{1}{2}(F1_{ref} + F1_{ans})$$

Penalizing **incorrect refusals** and **non-refusals**

→ **balanced evaluation** of the model’s over and under responsiveness

Calibrated Exact Match EM_{AC}^{F1}

$$AC^q = \frac{|A_G \cap A_D \cap A_R|}{|A_G \cap A_R|}$$

- Which generated claim (in A_R) is correct (in gold set A_G) and supported by docs (in A_D)
- Calibration: Disregards the claims that cannot be inferred from D

For the whole dataset with multiple questions {q1 ...qk},

$$AC = \frac{1}{k} \sum_{q_i \in A_g \cap A_r} AC^{q_i}$$

Calibrated Exact Match EM_{AC}^{F1}

$$AC = \frac{1}{k} \sum_{q_i \in A_g \cap A_r} AC^{q_i}$$

if $k = |A_r|$ P_{AC}

Precision oriented

if $k = |A_g|$ R_{AC}

Recall oriented

$$F1_{AC} = \frac{2P_{AC} \cdot R_{AC}}{P_{AC} + R_{AC}}$$

✅ Address imbalanced answerable/unanswerable class exploit

✅ Penalize models for relying solely on their pre-trained knowledge

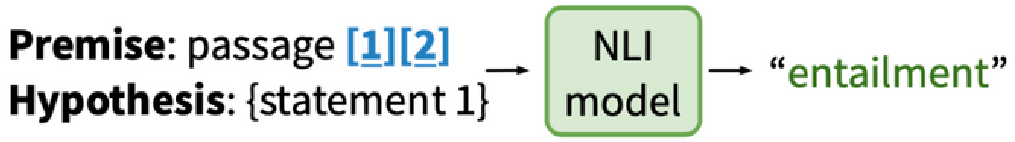
✅ Reward model for grounding answers on the provided documents

Citation Grounded F1 $F1_{CG}$

For a given statement si, **statement-wise citation recall (CR)** is computed by

$$R^{s_i}_{cite} = \phi(\{c_{i,1,\ldots,i,j}\}, s_i) \quad \rightarrow \{0, 1\}$$

Does the set of citations support statement si?

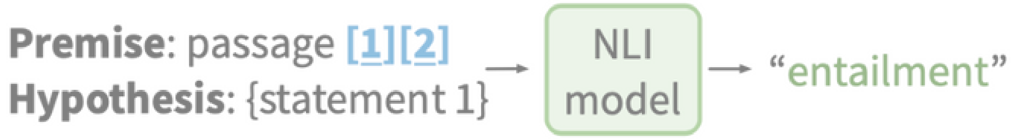


Citation Grounded F1 $F1_{CG}$

For a given statement si, **statement-wise citation recall (CR)** is computed by

$$R^{s_i}_{cite} = \phi(\{c_{i,1,\ldots,i,j}\}, s_i) \quad \rightarrow \{0, 1\}$$

Does the set of citations support statement si?



Premise: passage [1][2]
Hypothesis: {statement 1}

For a given citation ci, **citation precision (CP)** is computed by

$$R^{c_j}_{cite} = \phi(c_{i,j}, s_i)$$
$$\text{OR } \neg\phi(\{c_i, k | k \neq j\}, s_i)$$

(1) Does citation ci,j fully support statement si?

(2) Is the set of citations without ci,j insufficient to support statement si?

Citation Grounded F1 $F1_{CG}$

$$R_{cite} = \frac{1}{|A_r|} \sum_{C \in A_r^c} \frac{1}{|C|} \sum_{c_j \in C} R_{cite}^{s_i} \qquad P_{cite} = \frac{1}{|A_r|} \sum_{C \in A_r^c} \frac{1}{|C|} \sum_{c_j \in C} P_{cite}^{c_j}$$

$$\mathrm{F1}_{\mathrm{CG}} = \frac{2P_{cite} \cdot R_{cite}}{P_{cite} + R_{cite}}$$

A_r	Set of questions where model provided an answer
S	Set of statements in a generated response
C	Set of citations in a generated response
A_r^s	Set of responses (only statements, no citations) in the dataset
A_r^c	Set of responses (only citations, no statements) in the dataset

Trust-Score

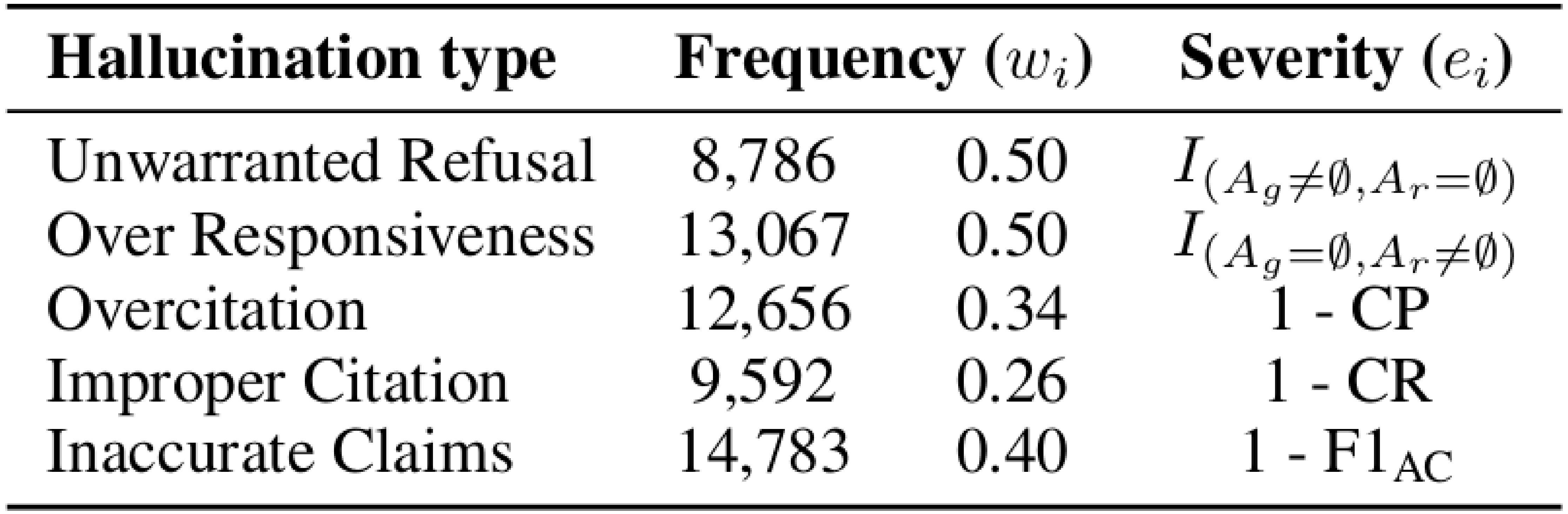
$$\text{TRUST-SCORE} = \frac{1}{3}(F1_{RG} + F1_{AC} + F1_{CG})$$

✅ Single trustworthiness score → rank models based on their trustworthiness.

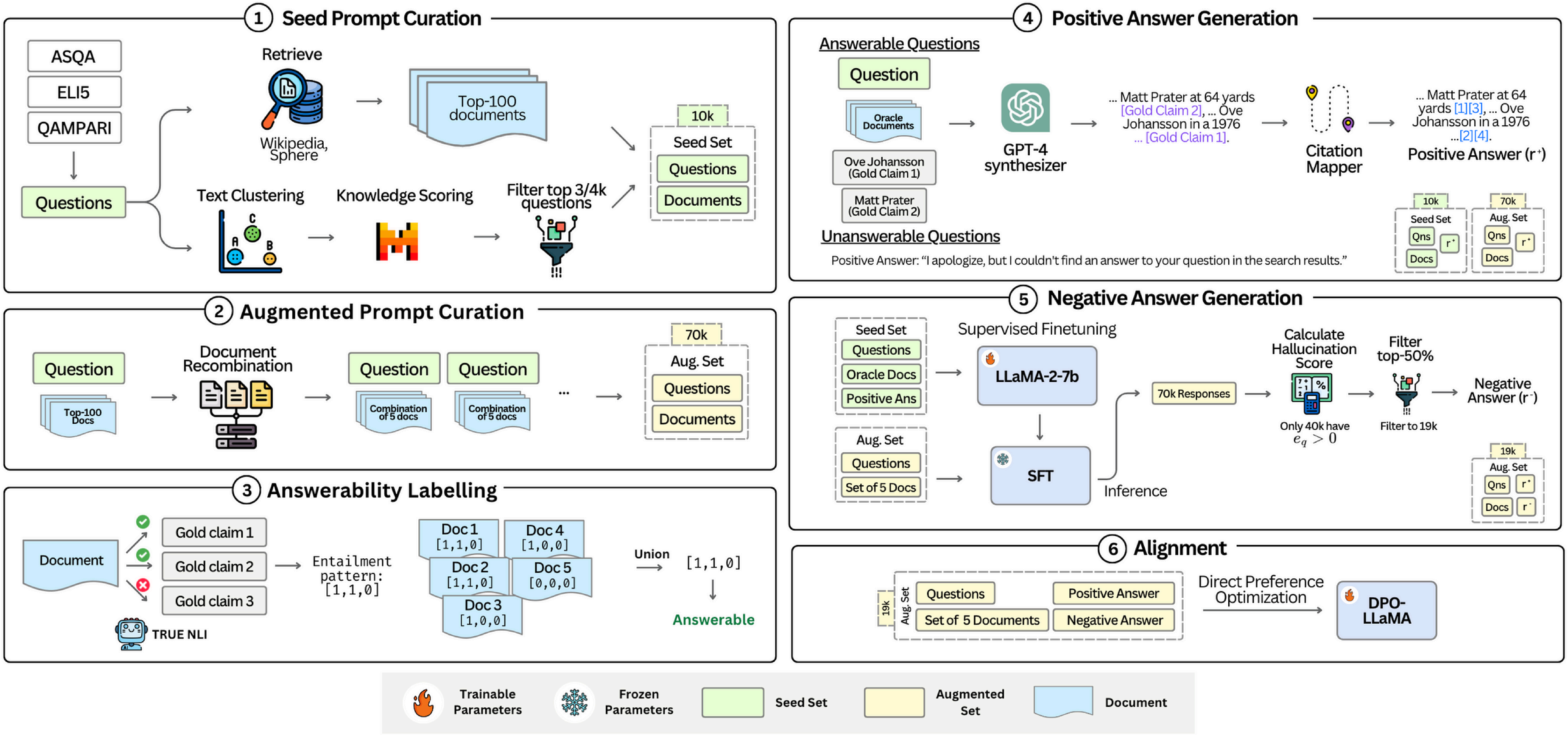
Trust-Align

An **alignment dataset** comprising **19K** questions, documents, and paired positive and negative responses, selected from the **top severity** of 40K hallucinations to enhance the groundedness of LLMs

Dataset covers **5 types of LLM hallucination** (opposite of LLM groundedness):

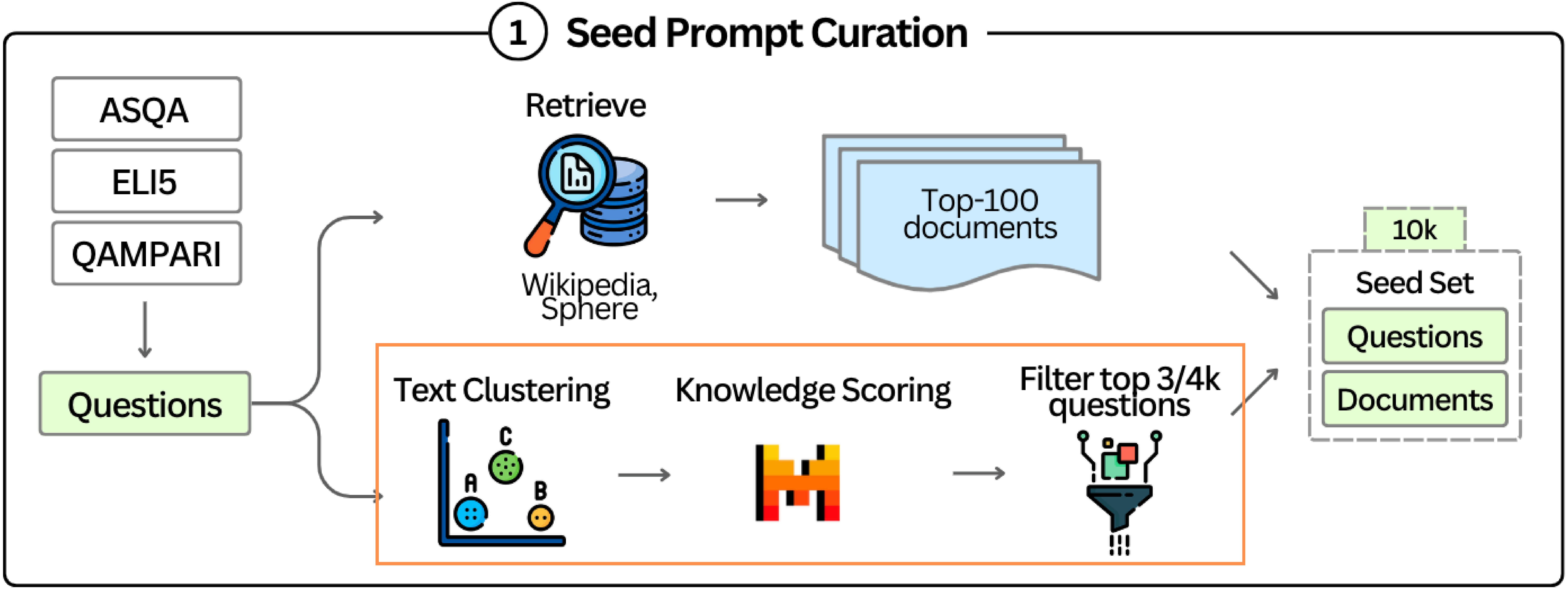


Hallucination type	**Frequency** (w_i)		**Severity** (e_i)
Unwarranted Refusal	8,786	0.50	$I_{(A_g \neq \emptyset, A_r = \emptyset)}$
Over Responsiveness	13,067	0.50	$I_{(A_g = \emptyset, A_r \neq \emptyset)}$
Overcitation	12,656	0.34	1 - CP
Improper Citation	9,592	0.26	1 - CR
Inaccurate Claims	14,783	0.40	1 - F1_{AC}

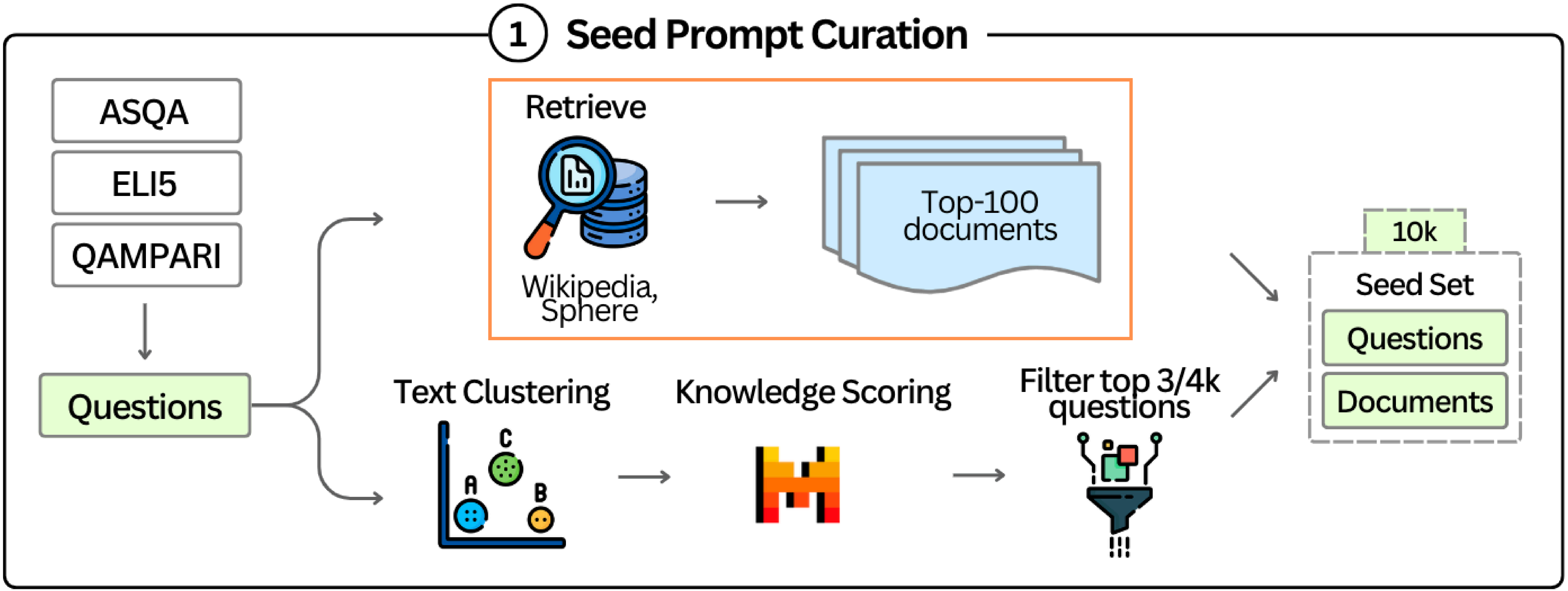


1 Seed Prompt Curation
ASQA
ELI5
QAMPARI
Questions
Retrieve
Wikipedia, Sphere
Top-100 documents
Text Clustering
C
A
B
Knowledge Scoring
Filter top 3/4k questions
10k
Seed Set
Questions
Documents
2 Augmented Prompt Curation
Question
Top-100 Docs
Document Recombination
Question
Combination of 5 docs
Question
Combination of 5 docs
...
70k
Aug. Set
Questions
Documents
3 Answerability Labelling
Document
Gold claim 1
Gold claim 2
Gold claim 3
TRUE NLI
Entailment pattern: [1,1,0]
Doc 1 [1,1,0]
Doc 4 [1,0,0]
Doc 2 [1,1,0]
Doc 5 [0,0,0]
Doc 3 [1,0,0]
Union
[1,1,0]
Answerable
4 Positive Answer Generation
Answerable Questions
Question
Oracle Documents
Ove Johansson (Gold Claim 1)
Matt Prater (Gold Claim 2)
GPT-4 synthesizer
... Matt Prater at 64 yards [Gold Claim 2], ... Ove Johansson in a 1976 ... [Gold Claim 1].
Citation Mapper
... Matt Prater at 64 yards [1][3], ... Ove Johansson in a 1976 ...[2][4].
Positive Answer (r⁺)
10k
Seed Set
Qns
r⁺
Docs
70k
Aug. Set
Qns
r⁺
Docs
Unanswerable Questions
Positive Answer: "I apologize, but I couldn't find an answer to your question in the search results."
5 Negative Answer Generation
Seed Set
Questions
Oracle Docs
Positive Ans
Supervised Finetuning
LLaMA-2-7b
Aug. Set
Questions
Set of 5 Docs
SFT
Inference
70k Responses
Calculate Hallucination Score
Only 40k have $e_q > 0$
Filter top-50%
Filter to 19k
Negative Answer (r⁻)
19k
Aug. Set
Qns
r⁺
Docs
r⁻
6 Alignment
19k
Aug. Set
Questions
Set of 5 Documents
Positive Answer
Negative Answer
Direct Preference Optimization
DPO-LLaMA
Trainable Parameters
Frozen Parameters
Seed Set
Augmented Set
Document

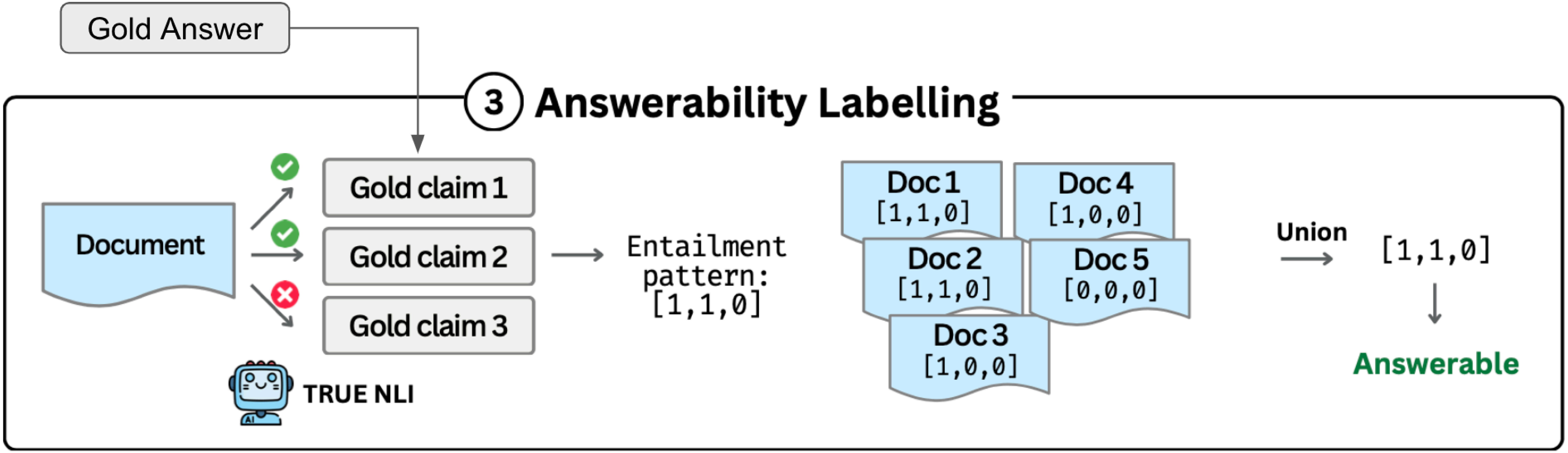
Select Quality Questions



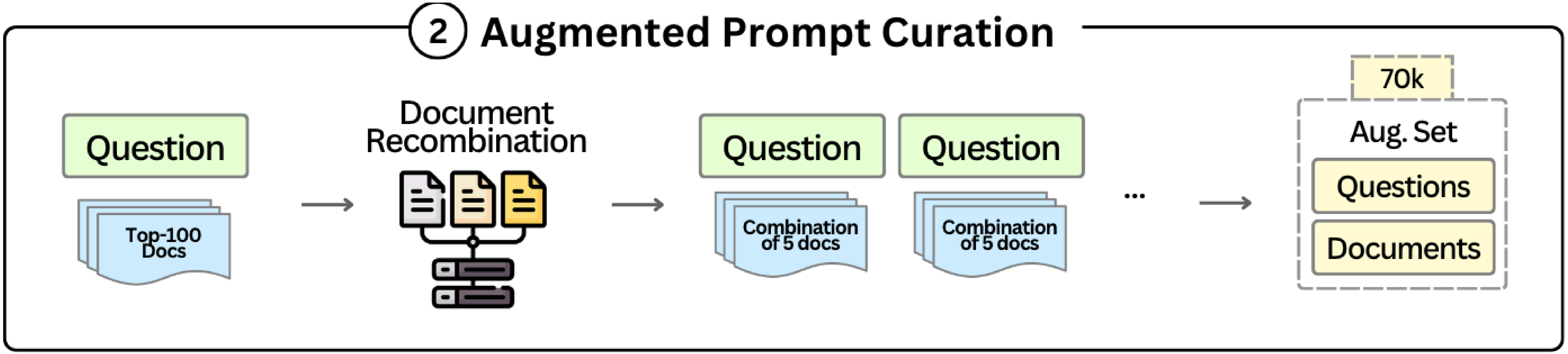
Collect Relevant Documents



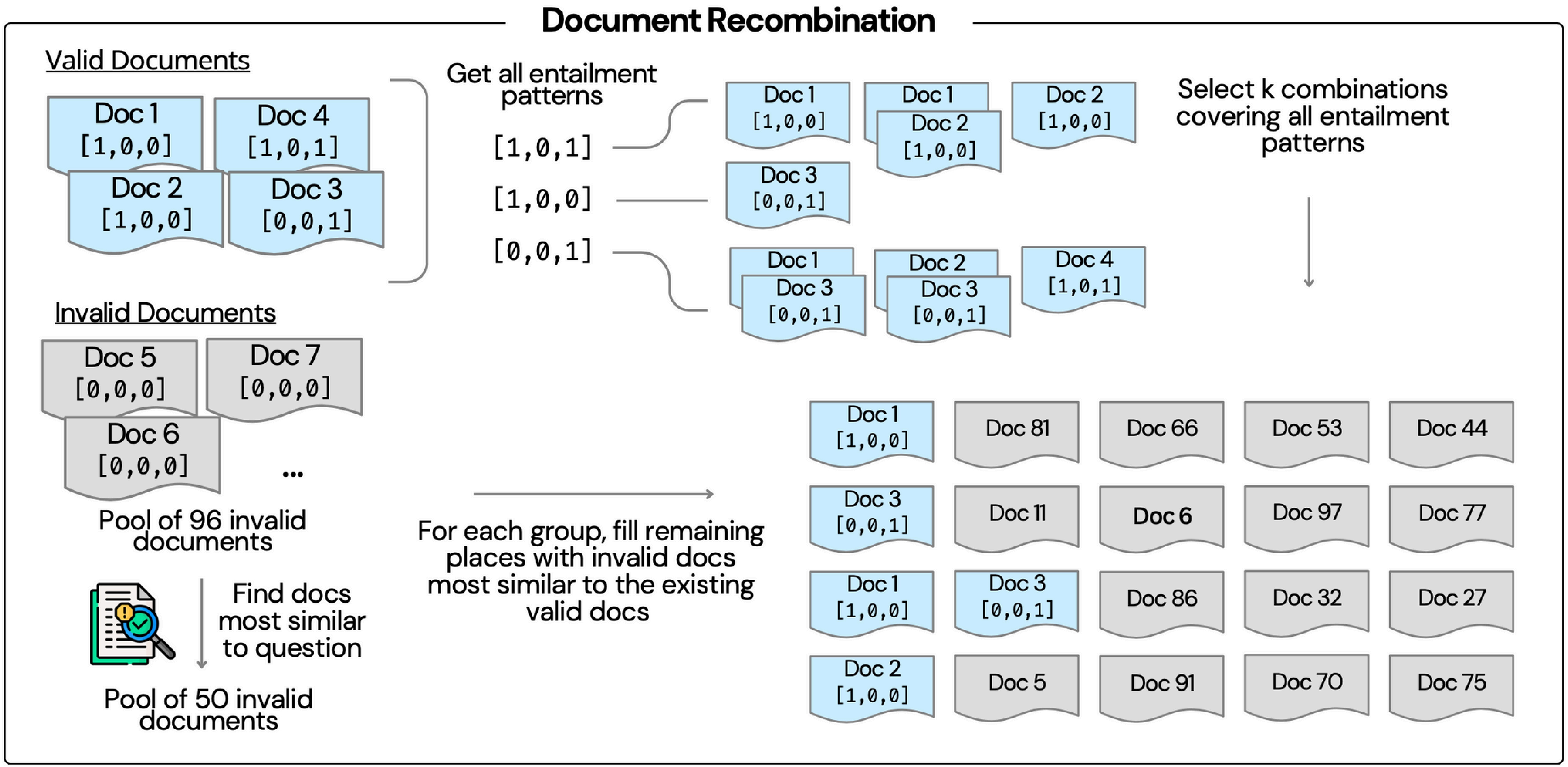
Derive the entailment pattern for each document



Diversify samples to trigger multiple hallucinations



Augmenting *(question, Documents)* pairs



Document Recombination

Get all entailment patterns

Select k combinations covering all entailment patterns

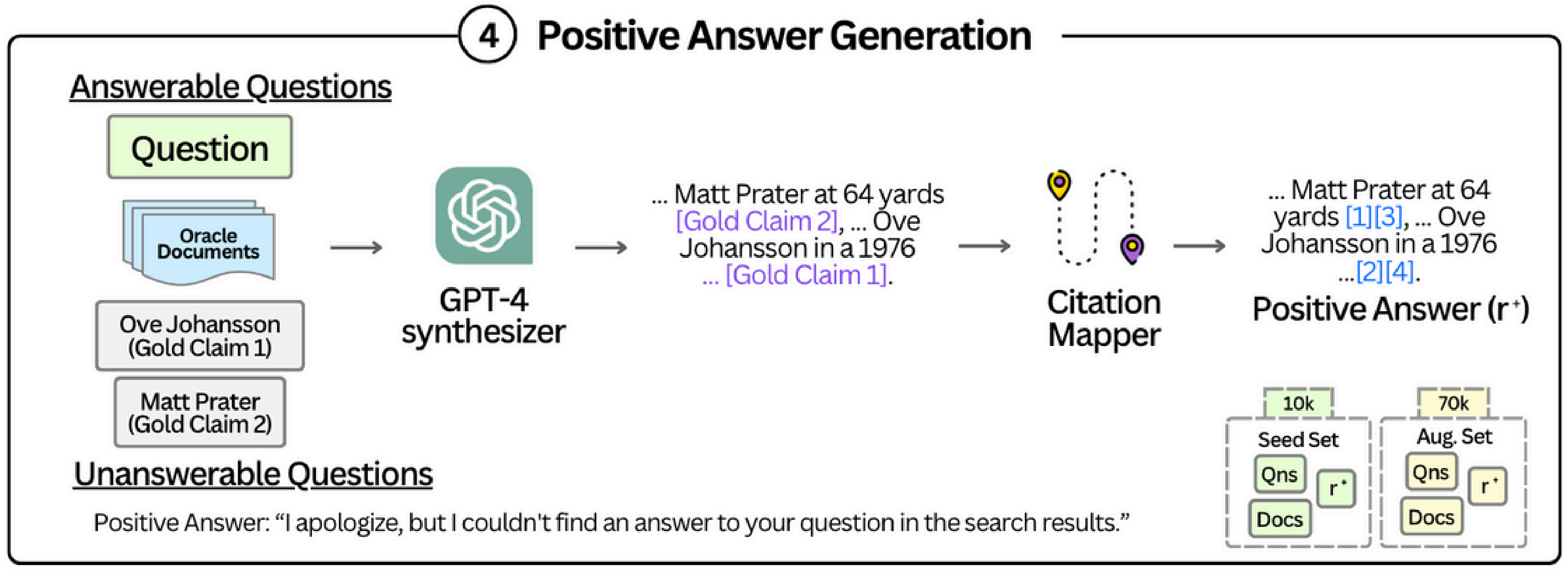
Pool of 96 invalid documents

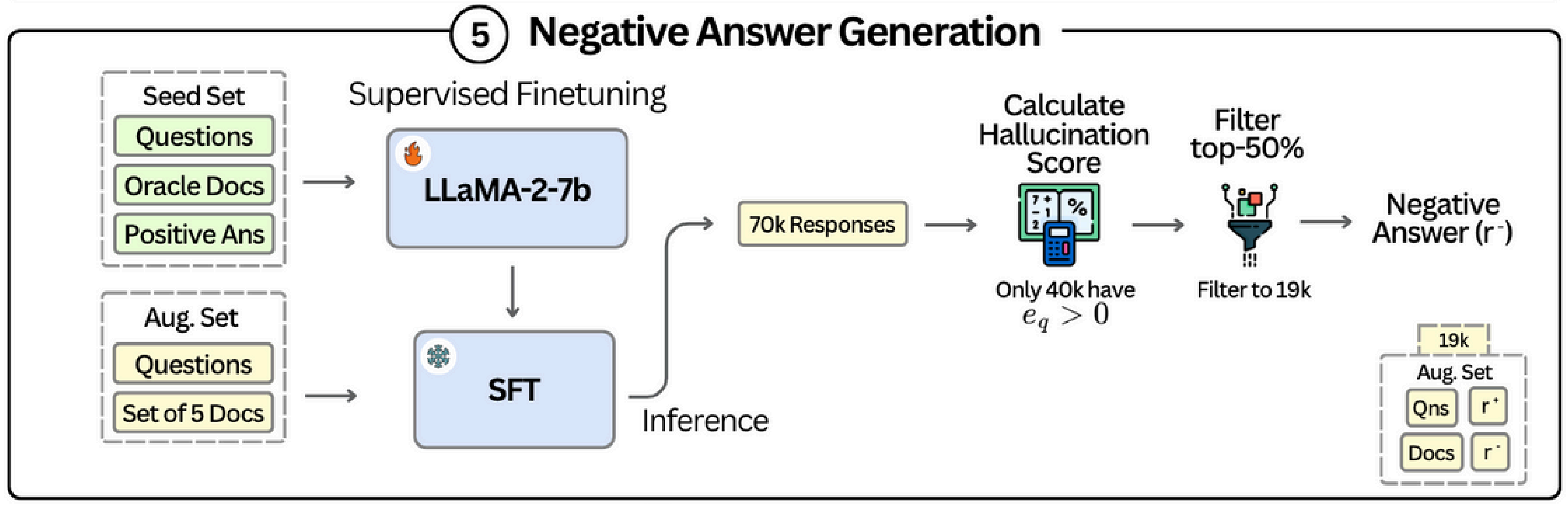
Find docs most similar to question

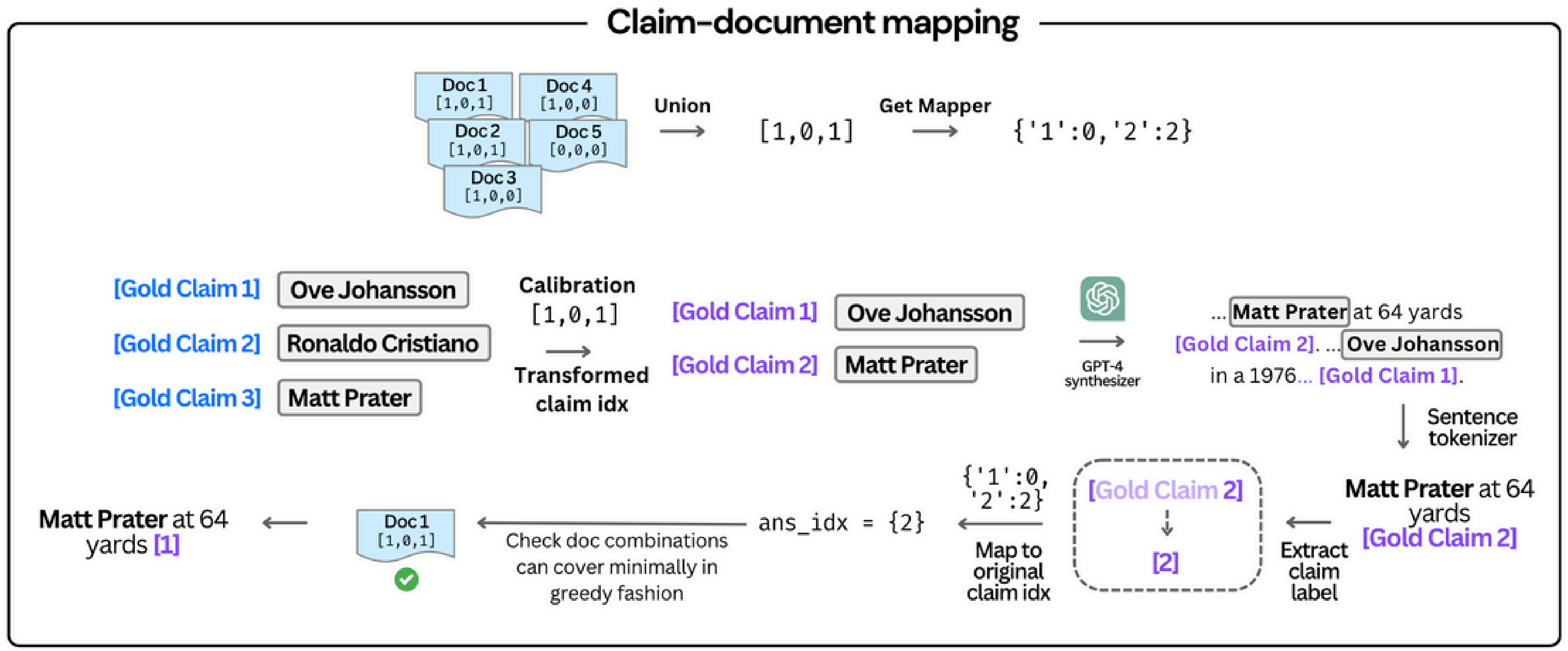
Pool of 50 invalid documents

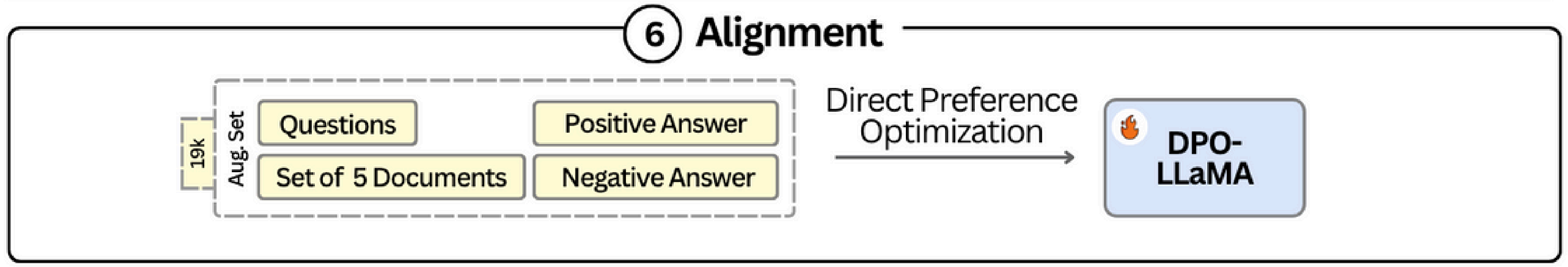
For each group, fill remaining places with invalid docs most similar to the existing valid docs

Obtaining Positive and Negative Answers

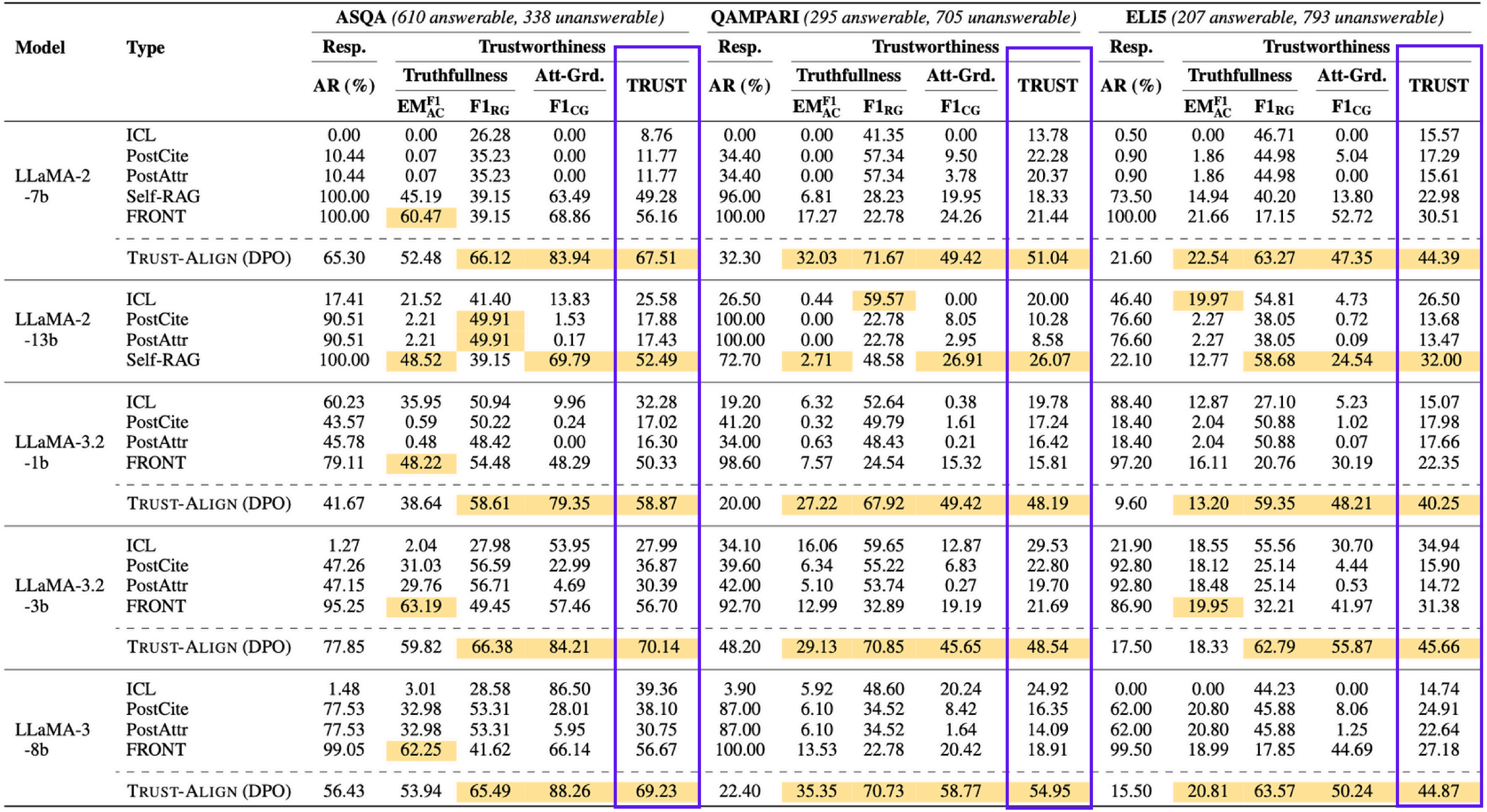

4 Positive Answer Generation
Answerable Questions
Question
Oracle Documents
Ove Johansson (Gold Claim 1)
Matt Prater (Gold Claim 2)
GPT-4 synthesizer
... Matt Prater at 64 yards [Gold Claim 2], ... Ove Johansson in a 1976 ... [Gold Claim 1].
Citation Mapper
... Matt Prater at 64 yards [1][3], ... Ove Johansson in a 1976 ...[2][4].
Positive Answer (r⁺)
10k
Seed Set
Qns
r⁺
Docs
70k
Aug. Set
Qns
r⁺
Docs
Unanswerable Questions
Positive Answer: "I apologize, but I couldn't find an answer to your question in the search results."




Claim-document mapping
Doc 1 [1,0,1]
Doc 4 [1,0,0]
Doc 2 [1,0,1]
Doc 5 [0,0,0]
Doc 3 [1,0,0]
Union
[1,0,1]
Get Mapper
{'1':0,'2':2}
[Gold Claim 1] Ove Johansson
[Gold Claim 2] Ronaldo Cristiano
[Gold Claim 3] Matt Prater
Calibration
[1,0,1]
Transformed claim idx
[Gold Claim 1] Ove Johansson
[Gold Claim 2] Matt Prater
GPT-4 synthesizer
... Matt Prater at 64 yards [Gold Claim 2]. ... Ove Johansson in a 1976... [Gold Claim 1].
Sentence tokenizer
Matt Prater at 64 yards [Gold Claim 2]
Extract claim label
[Gold Claim 2]
[2]
Map to original claim idx
{'1':0, '2':2}
ans_idx = {2}
Check doc combinations can cover minimally in greedy fashion
Doc 1 [1,0,1]
Matt Prater at 64 yards [1]

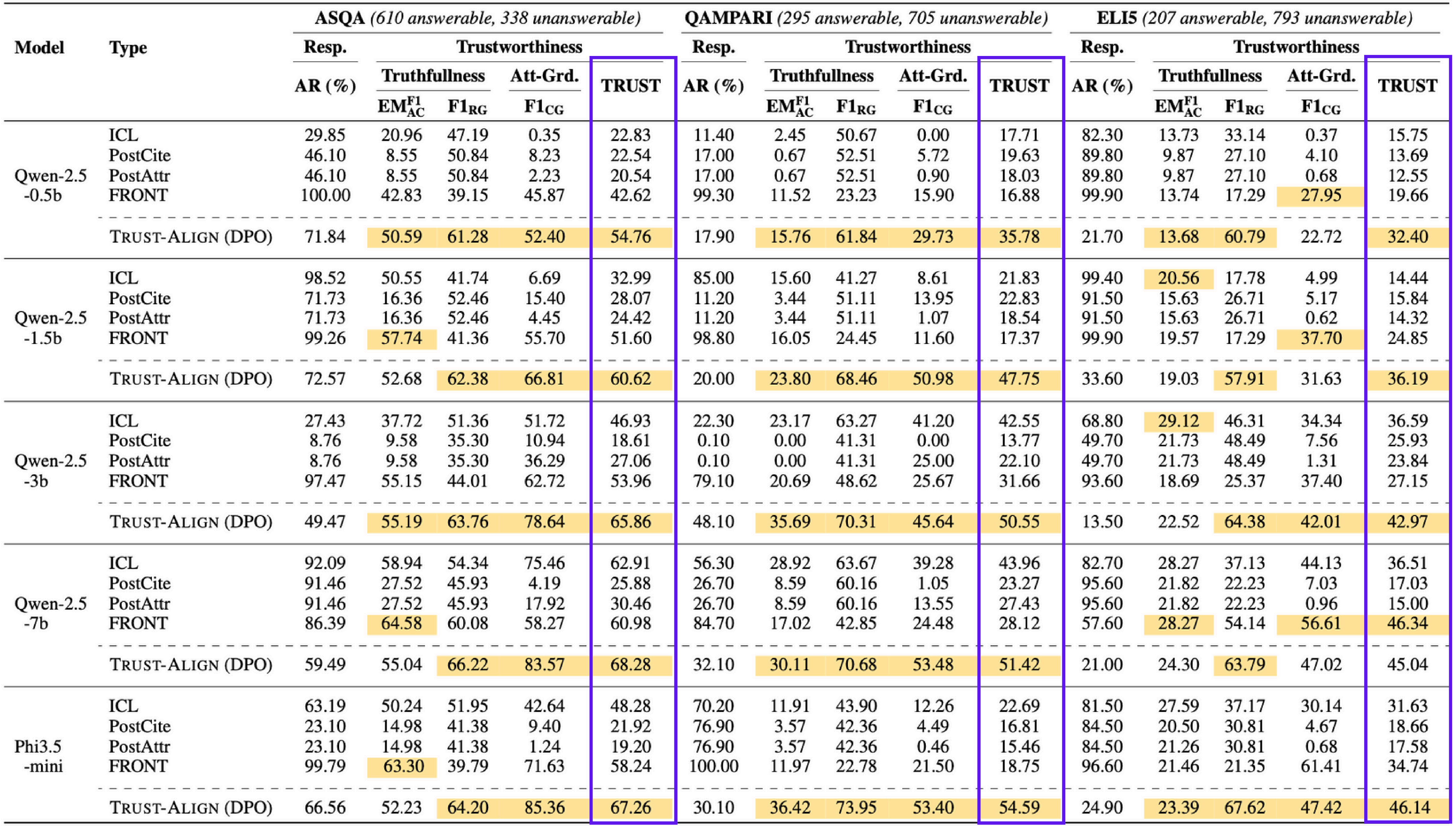


Trust-Align boosts trustworthiness over baselines



Model	Type	**ASQA** *(610 answerable, 338 unanswerable)*					**QAMPARI** *(295 answerable, 705 unanswerable)*					**ELI5** *(207 answerable, 793 unanswerable)*				
		Resp.	**Trustworthiness**				**Resp.**	**Trustworthiness**				**Resp.**	**Trustworthiness**			
		AR (%)	**Truthfullness**		**Att-Grd.**	**TRUST**	**AR (%)**	**Truthfullness**		**Att-Grd.**	**TRUST**	**AR (%)**	**Truthfullness**		**Att-Grd.**	**TRUST**
			$\mathbf{EM}^{\mathbf{F1}}_{\mathbf{AC}}$	$\mathbf{F1}_{\mathbf{RG}}$	$\mathbf{F1}_{\mathbf{CG}}$			$\mathbf{EM}^{\mathbf{F1}}_{\mathbf{AC}}$	$\mathbf{F1}_{\mathbf{RG}}$	$\mathbf{F1}_{\mathbf{CG}}$			$\mathbf{EM}^{\mathbf{F1}}_{\mathbf{AC}}$	$\mathbf{F1}_{\mathbf{RG}}$	$\mathbf{F1}_{\mathbf{CG}}$	
LLaMA-2 -7b	ICL	0.00	0.00	26.28	0.00	8.76	0.00	0.00	41.35	0.00	13.78	0.50	0.00	46.71	0.00	15.57
	PostCite	10.44	0.07	35.23	0.00	11.77	34.40	0.00	57.34	9.50	22.28	0.90	1.86	44.98	5.04	17.29
	PostAttr	10.44	0.07	35.23	0.00	11.77	34.40	0.00	57.34	3.78	20.37	0.90	1.86	44.98	0.00	15.61
	Self-RAG	100.00	45.19	39.15	63.49	49.28	96.00	6.81	28.23	19.95	18.33	73.50	14.94	40.20	13.80	22.98
	FRONT	100.00	60.47	39.15	68.86	56.16	100.00	17.27	22.78	24.26	21.44	100.00	21.66	17.15	52.72	30.51
	TRUST-ALIGN (DPO)	65.30	52.48	66.12	83.94	67.51	32.30	32.03	71.67	49.42	51.04	21.60	22.54	63.27	47.35	44.39
LLaMA-2 -13b	ICL	17.41	21.52	41.40	13.83	25.58	26.50	0.44	59.57	0.00	20.00	46.40	19.97	54.81	4.73	26.50
	PostCite	90.51	2.21	49.91	1.53	17.88	100.00	0.00	22.78	8.05	10.28	76.60	2.27	38.05	0.72	13.68
	PostAttr	90.51	2.21	49.91	0.17	17.43	100.00	0.00	22.78	2.95	8.58	76.60	2.27	38.05	0.09	13.47
	Self-RAG	100.00	48.52	39.15	69.79	52.49	72.70	2.71	48.58	26.91	26.07	22.10	12.77	58.68	24.54	32.00
LLaMA-3.2 -1b	ICL	60.23	35.95	50.94	9.96	32.28	19.20	6.32	52.64	0.38	19.78	88.40	12.87	27.10	5.23	15.07
	PostCite	43.57	0.59	50.22	0.24	17.02	41.20	0.32	49.79	1.61	17.24	18.40	2.04	50.88	1.02	17.98
	PostAttr	45.78	0.48	48.42	0.00	16.30	34.00	0.63	48.43	0.21	16.42	18.40	2.04	50.88	0.07	17.66
	FRONT	79.11	48.22	54.48	48.29	50.33	98.60	7.57	24.54	15.32	15.81	97.20	16.11	20.76	30.19	22.35
	TRUST-ALIGN (DPO)	41.67	38.64	58.61	79.35	58.87	20.00	27.22	67.92	49.42	48.19	9.60	13.20	59.35	48.21	40.25
LLaMA-3.2 -3b	ICL	1.27	2.04	27.98	53.95	27.99	34.10	16.06	59.65	12.87	29.53	21.90	18.55	55.56	30.70	34.94
	PostCite	47.26	31.03	56.59	22.99	36.87	39.60	6.34	55.22	6.83	22.80	92.80	18.12	25.14	4.44	15.90
	PostAttr	47.15	29.76	56.71	4.69	30.39	42.00	5.10	53.74	0.27	19.70	92.80	18.48	25.14	0.53	14.72
	FRONT	95.25	63.19	49.45	57.46	56.70	92.70	12.99	32.89	19.19	21.69	86.90	19.95	32.21	41.97	31.38
	TRUST-ALIGN (DPO)	77.85	59.82	66.38	84.21	70.14	48.20	29.13	70.85	45.65	48.54	17.50	18.33	62.79	55.87	45.66
LLaMA-3 -8b	ICL	1.48	3.01	28.58	86.50	39.36	3.90	5.92	48.60	20.24	24.92	0.00	0.00	44.23	0.00	14.74
	PostCite	77.53	32.98	53.31	28.01	38.10	87.00	6.10	34.52	8.42	16.35	62.00	20.80	45.88	8.06	24.91
	PostAttr	77.53	32.98	53.31	5.95	30.75	87.00	6.10	34.52	1.64	14.09	62.00	20.80	45.88	1.25	22.64
	FRONT	99.05	62.25	41.62	66.14	56.67	100.00	13.53	22.78	20.42	18.91	99.50	18.99	17.85	44.69	27.18
	TRUST-ALIGN (DPO)	56.43	53.94	65.49	88.26	69.23	22.40	35.35	70.73	58.77	54.95	15.50	20.81	63.57	50.24	44.87

Trust-Align generalizes across model families and sizes



Model	Type	**ASQA** *(610 answerable, 338 unanswerable)*					**QAMPARI** *(295 answerable, 705 unanswerable)*					**ELI5** *(207 answerable, 793 unanswerable)*				
		Resp.	**Trustworthiness**				**Resp.**	**Trustworthiness**				**Resp.**	**Trustworthiness**			
		AR (%)	**Truthfullness**		**Att-Grd.**	**TRUST**	**AR (%)**	**Truthfullness**		**Att-Grd.**	**TRUST**	**AR (%)**	**Truthfullness**		**Att-Grd.**	**TRUST**
			$\text{EM}^{\text{F1}}_{\text{AC}}$	**F1_{RG}**	**F1_{CG}**			**$\text{EM}^{\text{F1}}_{\text{AC}}$**	**F1_{RG}**	**F1_{CG}**			**$\text{EM}^{\text{F1}}_{\text{AC}}$**	**F1_{RG}**	**F1_{CG}**	
Qwen-2.5 -0.5b	ICL	29.85	20.96	47.19	0.35	22.83	11.40	2.45	50.67	0.00	17.71	82.30	13.73	33.14	0.37	15.75
	PostCite	46.10	8.55	50.84	8.23	22.54	17.00	0.67	52.51	5.72	19.63	89.80	9.87	27.10	4.10	13.69
	PostAttr	46.10	8.55	50.84	2.23	20.54	17.00	0.67	52.51	0.90	18.03	89.80	9.87	27.10	0.68	12.55
	FRONT	100.00	42.83	39.15	45.87	42.62	99.30	11.52	23.23	15.90	16.88	99.90	13.74	17.29	27.95	19.66
	TRUST-ALIGN (DPO)	71.84	50.59	61.28	52.40	54.76	17.90	15.76	61.84	29.73	35.78	21.70	13.68	60.79	22.72	32.40
Qwen-2.5 -1.5b	ICL	98.52	50.55	41.74	6.69	32.99	85.00	15.60	41.27	8.61	21.83	99.40	20.56	17.78	4.99	14.44
	PostCite	71.73	16.36	52.46	15.40	28.07	11.20	3.44	51.11	13.95	22.83	91.50	15.63	26.71	5.17	15.84
	PostAttr	71.73	16.36	52.46	4.45	24.42	11.20	3.44	51.11	1.07	18.54	91.50	15.63	26.71	0.62	14.32
	FRONT	99.26	57.74	41.36	55.70	51.60	98.80	16.05	24.45	11.60	17.37	99.90	19.57	17.29	37.70	24.85
	TRUST-ALIGN (DPO)	72.57	52.68	62.38	66.81	60.62	20.00	23.80	68.46	50.98	47.75	33.60	19.03	57.91	31.63	36.19
Qwen-2.5 -3b	ICL	27.43	37.72	51.36	51.72	46.93	22.30	23.17	63.27	41.20	42.55	68.80	29.12	46.31	34.34	36.59
	PostCite	8.76	9.58	35.30	10.94	18.61	0.10	0.00	41.31	0.00	13.77	49.70	21.73	48.49	7.56	25.93
	PostAttr	8.76	9.58	35.30	36.29	27.06	0.10	0.00	41.31	25.00	22.10	49.70	21.73	48.49	1.31	23.84
	FRONT	97.47	55.15	44.01	62.72	53.96	79.10	20.69	48.62	25.67	31.66	93.60	18.69	25.37	37.40	27.15
	TRUST-ALIGN (DPO)	49.47	55.19	63.76	78.64	65.86	48.10	35.69	70.31	45.64	50.55	13.50	22.52	64.38	42.01	42.97
Qwen-2.5 -7b	ICL	92.09	58.94	54.34	75.46	62.91	56.30	28.92	63.67	39.28	43.96	82.70	28.27	37.13	44.13	36.51
	PostCite	91.46	27.52	45.93	4.19	25.88	26.70	8.59	60.16	1.05	23.27	95.60	21.82	22.23	7.03	17.03
	PostAttr	91.46	27.52	45.93	17.92	30.46	26.70	8.59	60.16	13.55	27.43	95.60	21.82	22.23	0.96	15.00
	FRONT	86.39	64.58	60.08	58.27	60.98	84.70	17.02	42.85	24.48	28.12	57.60	28.27	54.14	56.61	46.34
	TRUST-ALIGN (DPO)	59.49	55.04	66.22	83.57	68.28	32.10	30.11	70.68	53.48	51.42	21.00	24.30	63.79	47.02	45.04
Phi3.5 -mini	ICL	63.19	50.24	51.95	42.64	48.28	70.20	11.91	43.90	12.26	22.69	81.50	27.59	37.17	30.14	31.63
	PostCite	23.10	14.98	41.38	9.40	21.92	76.90	3.57	42.36	4.49	16.81	84.50	20.50	30.81	4.67	18.66
	PostAttr	23.10	14.98	41.38	1.24	19.20	76.90	3.57	42.36	0.46	15.46	84.50	21.26	30.81	0.68	17.58
	FRONT	99.79	63.30	39.79	71.63	58.24	100.00	11.97	22.78	21.50	18.75	96.60	21.46	21.35	61.41	34.74
	TRUST-ALIGN (DPO)	66.56	52.23	64.20	85.36	67.26	30.10	36.42	73.95	53.40	54.59	24.90	23.39	67.62	47.42	46.14

SFT of GPT-4o on Trust-Align

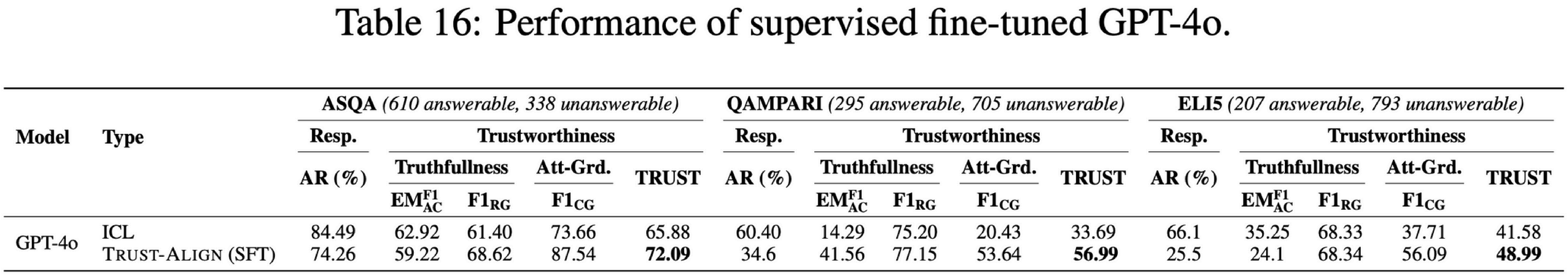


Table 16: Performance of supervised fine-tuned GPT-4o.

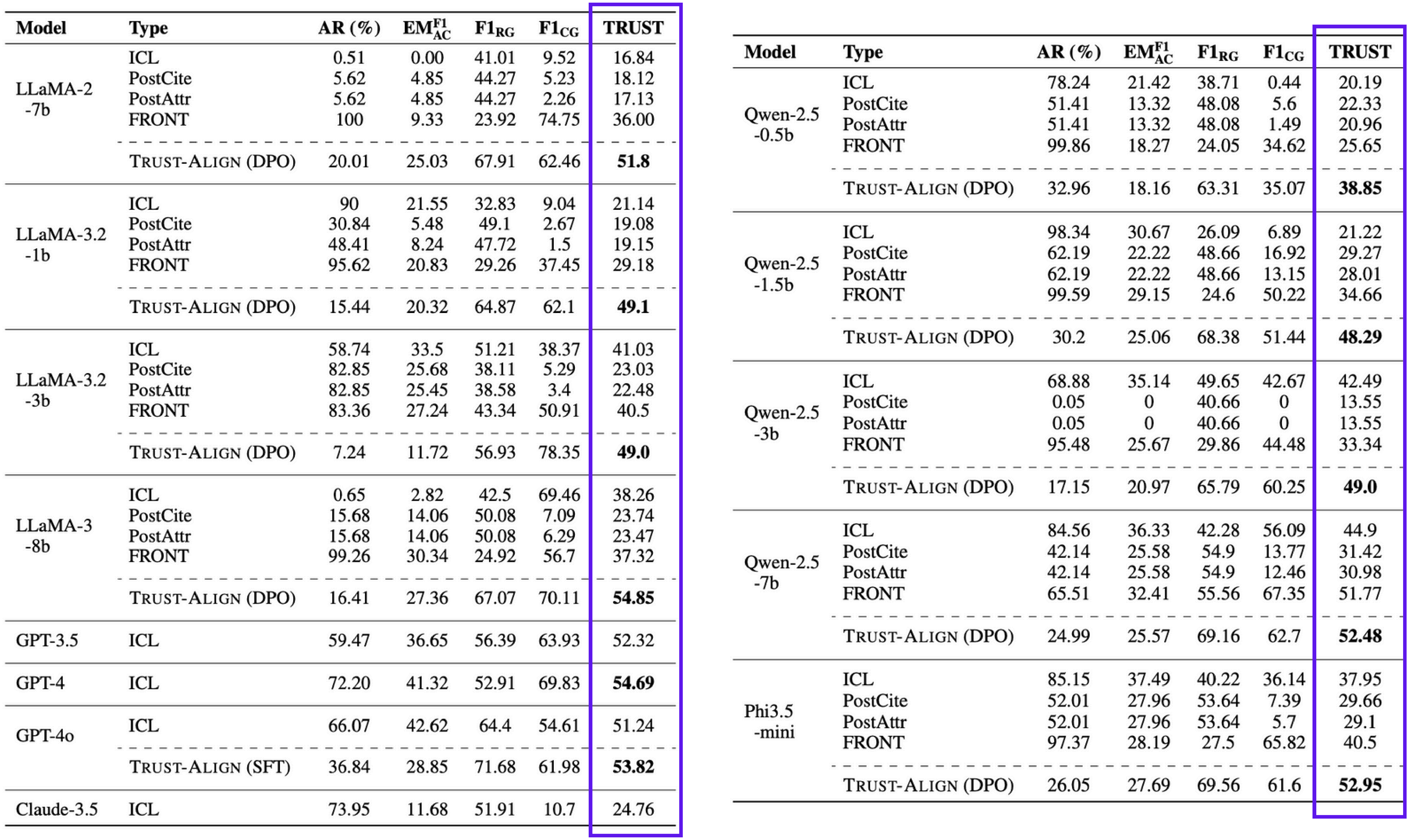
Model	**Type**	**ASQA** *(610 answerable, 338 unanswerable)*					**QAMPARI** *(295 answerable, 705 unanswerable)*					**ELI5** *(207 answerable, 793 unanswerable)*				
		Resp.	**Trustworthiness**				**Resp.**	**Trustworthiness**				**Resp.**	**Trustworthiness**			
		AR (%)	**Truthfullness**		**Att-Grd.**	**TRUST**	**AR (%)**	**Truthfullness**		**Att-Grd.**	**TRUST**	**AR (%)**	**Truthfullness**		**Att-Grd.**	**TRUST**
			$\mathbf{EM}^{\mathbf{F1}}_{\mathbf{AC}}$	$\mathbf{F1}_{\mathbf{RG}}$	$\mathbf{F1}_{\mathbf{CG}}$			$\mathbf{EM}^{\mathbf{F1}}_{\mathbf{AC}}$	$\mathbf{F1}_{\mathbf{RG}}$	$\mathbf{F1}_{\mathbf{CG}}$			$\mathbf{EM}^{\mathbf{F1}}_{\mathbf{AC}}$	$\mathbf{F1}_{\mathbf{RG}}$	$\mathbf{F1}_{\mathbf{CG}}$	
GPT-4o	ICL	84.49	62.92	61.40	73.66	65.88	60.40	14.29	75.20	20.43	33.69	66.1	35.25	68.33	37.71	41.58
	TRUST-ALIGN (SFT)	74.26	59.22	68.62	87.54	**72.09**	34.6	41.56	77.15	53.64	**56.99**	25.5	24.1	68.34	56.09	**48.99**

When aligned using a subset of Trust-Align data, GPT-4o improves its Trust Score by 6.21 (ASQA), 23.3 (QAMPARI), and 7.41 (ELI5) points.

=> observe the potential impact of such an alignment on flagship models

Improvements Generalizes on Out-of-Domain Data

Table 7: Generalization test results on ExpertQA using refusal prompting.



Model	Type	AR (%)	$\text{EM}^{\text{F1}}_{\text{AC}}$	F1_{RG}	F1_{CG}	TRUST
LLaMA-2 -7b	ICL	0.51	0.00	41.01	9.52	16.84
	PostCite	5.62	4.85	44.27	5.23	18.12
	PostAttr	5.62	4.85	44.27	2.26	17.13
	FRONT	100	9.33	23.92	74.75	36.00
	Trust-Align (DPO)	20.01	25.03	67.91	62.46	**51.8**
LLaMA-3.2 -1b	ICL	90	21.55	32.83	9.04	21.14
	PostCite	30.84	5.48	49.1	2.67	19.08
	PostAttr	48.41	8.24	47.72	1.5	19.15
	FRONT	95.62	20.83	29.26	37.45	29.18
	Trust-Align (DPO)	15.44	20.32	64.87	62.1	**49.1**
LLaMA-3.2 -3b	ICL	58.74	33.5	51.21	38.37	41.03
	PostCite	82.85	25.68	38.11	5.29	23.03
	PostAttr	82.85	25.45	38.58	3.4	22.48
	FRONT	83.36	27.24	43.34	50.91	40.5
	Trust-Align (DPO)	7.24	11.72	56.93	78.35	**49.0**
LLaMA-3 -8b	ICL	0.65	2.82	42.5	69.46	38.26
	PostCite	15.68	14.06	50.08	7.09	23.74
	PostAttr	15.68	14.06	50.08	6.29	23.47
	FRONT	99.26	30.34	24.92	56.7	37.32
	Trust-Align (DPO)	16.41	27.36	67.07	70.11	**54.85**
GPT-3.5	ICL	59.47	36.65	56.39	63.93	52.32
GPT-4	ICL	72.20	41.32	52.91	69.83	**54.69**
GPT-4o	ICL	66.07	42.62	64.4	54.61	51.24
	Trust-Align (SFT)	36.84	28.85	71.68	61.98	**53.82**
Claude-3.5	ICL	73.95	11.68	51.91	10.7	24.76

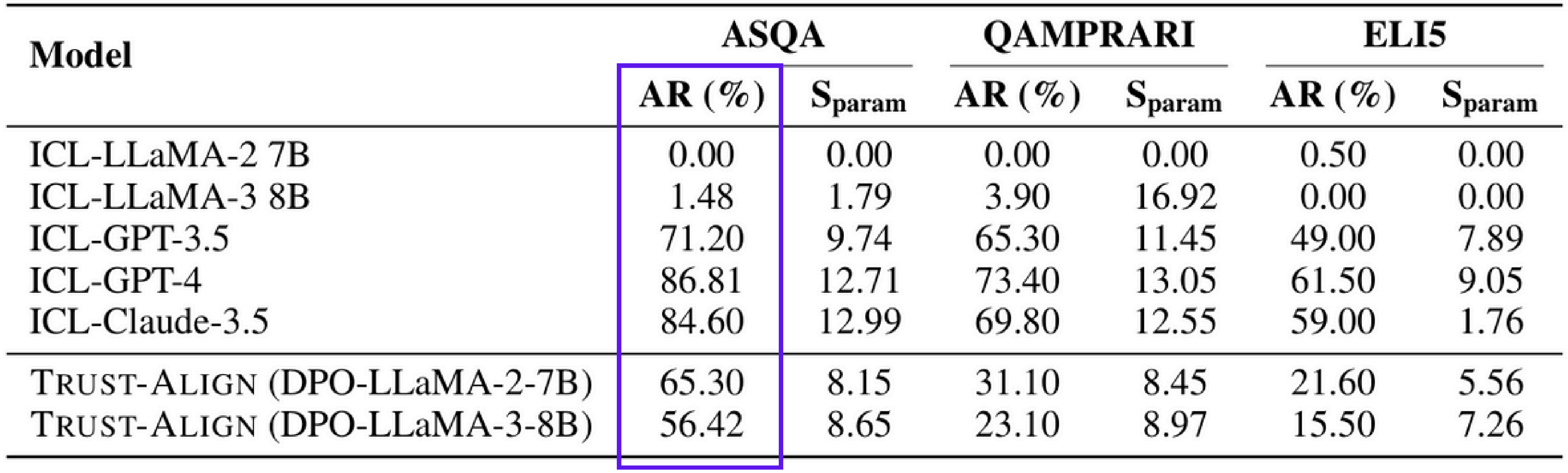
Model	Type	AR (%)	$\text{EM}^{\text{F1}}_{\text{AC}}$	F1_{RG}	F1_{CG}	TRUST
Qwen-2.5 -0.5b	ICL	78.24	21.42	38.71	0.44	20.19
	PostCite	51.41	13.32	48.08	5.6	22.33
	PostAttr	51.41	13.32	48.08	1.49	20.96
	FRONT	99.86	18.27	24.05	34.62	25.65
	Trust-Align (DPO)	32.96	18.16	63.31	35.07	**38.85**
Qwen-2.5 -1.5b	ICL	98.34	30.67	26.09	6.89	21.22
	PostCite	62.19	22.22	48.66	16.92	29.27
	PostAttr	62.19	22.22	48.66	13.15	28.01
	FRONT	99.59	29.15	24.6	50.22	34.66
	Trust-Align (DPO)	30.2	25.06	68.38	51.44	**48.29**
Qwen-2.5 -3b	ICL	68.88	35.14	49.65	42.67	42.49
	PostCite	0.05	0	40.66	0	13.55
	PostAttr	0.05	0	40.66	0	13.55
	FRONT	95.48	25.67	29.86	44.48	33.34
	Trust-Align (DPO)	17.15	20.97	65.79	60.25	**49.0**
Qwen-2.5 -7b	ICL	84.56	36.33	42.28	56.09	44.9
	PostCite	42.14	25.58	54.9	13.77	31.42
	PostAttr	42.14	25.58	54.9	12.46	30.98
	FRONT	65.51	32.41	55.56	67.35	51.77
	Trust-Align (DPO)	24.99	25.57	69.16	62.7	**52.48**
Phi3.5 -mini	ICL	85.15	37.49	40.22	36.14	37.95
	PostCite	52.01	27.96	53.64	7.39	29.66
	PostAttr	52.01	27.96	53.64	5.7	29.1
	FRONT	97.37	28.19	27.5	65.82	40.5
	Trust-Align (DPO)	26.05	27.69	69.56	61.6	**52.95**

Tendency of Grounding Knowledge on External Documents

$$\mathrm{S}_{\text{param}} = \frac{1}{|\mathcal{N}_r|} \sum_{q_i \in \mathcal{N}_r} \frac{|(A_R - (A_R \cap A_D)) \cap A_G|}{|A_R|}$$

Quantify the proportion of correctly generated claims for unanswerable questions

Table 10: Detection of parametric knowledge usage under refusal prompting.



Model	ASQA		QAMPRARI		ELI5	
	AR (%)	$\mathrm{S}_{\text{param}}$	AR (%)	$\mathrm{S}_{\text{param}}$	AR (%)	$\mathrm{S}_{\text{param}}$
ICL-LLaMA-2 7B	0.00	0.00	0.00	0.00	0.50	0.00
ICL-LLaMA-3 8B	1.48	1.79	3.90	16.92	0.00	0.00
ICL-GPT-3.5	71.20	9.74	65.30	11.45	49.00	7.89
ICL-GPT-4	86.81	12.71	73.40	13.05	61.50	9.05
ICL-Claude-3.5	84.60	12.99	69.80	12.55	59.00	1.76
TRUST-ALIGN (DPO-LLaMA-2-7B)	65.30	8.15	31.10	8.45	21.60	5.56
TRUST-ALIGN (DPO-LLaMA-3-8B)	56.42	8.65	23.10	8.97	15.50	7.26

Responsive models tend to rely on parametric knowledge more frequently.

Primary Sources of LLM Erroneous Generation

Sources of errors for answerable questions:

1. Parametric knowledge-based hallucination

2. Information extraction failures

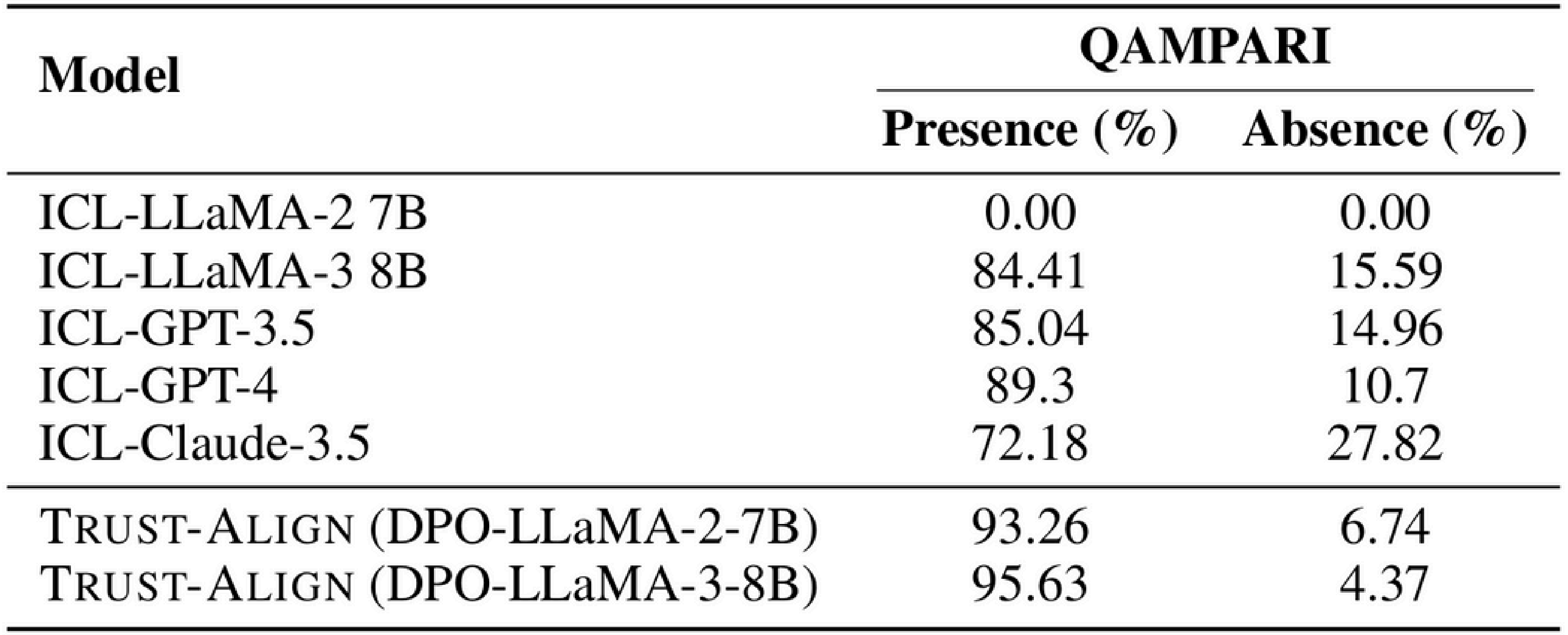
$$\text{Presence} = \frac{1}{|\mathcal{N}_e|} \sum_{q_i \in \mathcal{A}_e} \frac{|A_R^e \cap A_D|}{|A_R^e|}$$

$$\text{Absence} = \frac{1}{|\mathcal{N}_e|} \sum_{q_i \in \mathcal{A}_e} \frac{|A_R^e - (A_R^e \cap A_D)|}{|A_R^e|}$$

A higher tendency to produce erroneous answers based on their parametric knowledge

↓

More susceptible to hallucinations stemming from their parametric knowledge



Model	QAMPARI	
	Presence (%)	Absence (%)
ICL-LLaMA-2 7B	0.00	0.00
ICL-LLaMA-3 8B	84.41	15.59
ICL-GPT-3.5	85.04	14.96
ICL-GPT-4	89.3	10.7
ICL-Claude-3.5	72.18	27.82
TRUST-ALIGN (DPO-LLaMA-2-7B)	93.26	6.74
TRUST-ALIGN (DPO-LLaMA-3-8B)	95.63	4.37

Thank you!

📎 Paper

✏️ Codebase