

# **Measuring and Enhancing Trustworthiness of LLMs in RAG through Grounded Attributions and Learning to Refuse**

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

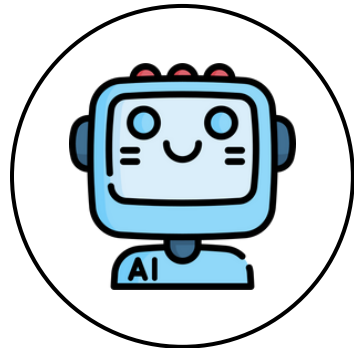
# What is a good quality output for RAG?

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.



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

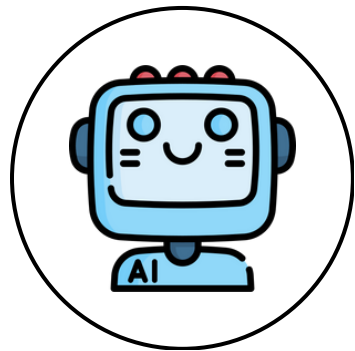
# What is a good quality output for RAG?

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.

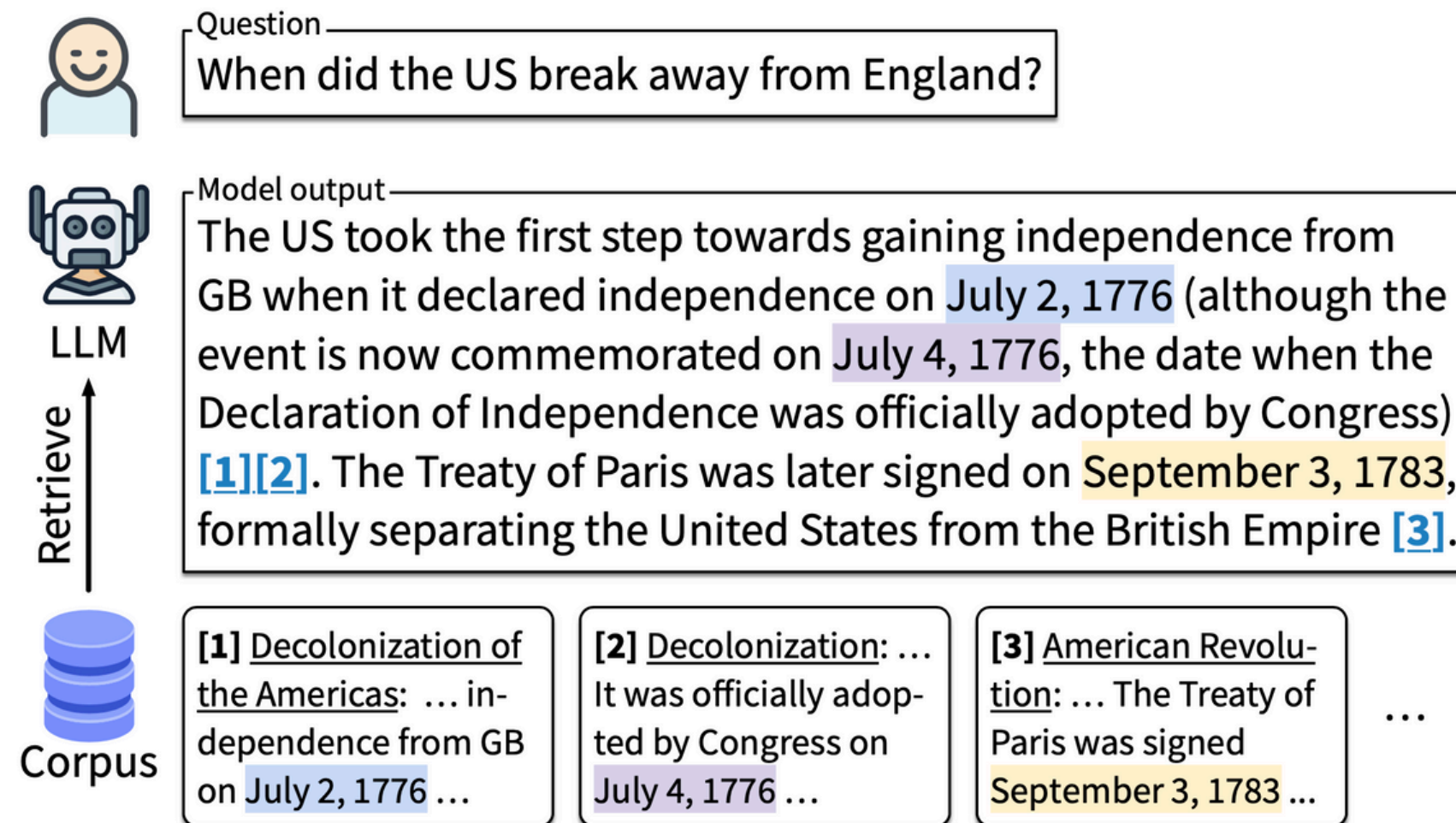


Barack Obama was born in Hawaii on August 4 1961 [1].

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

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

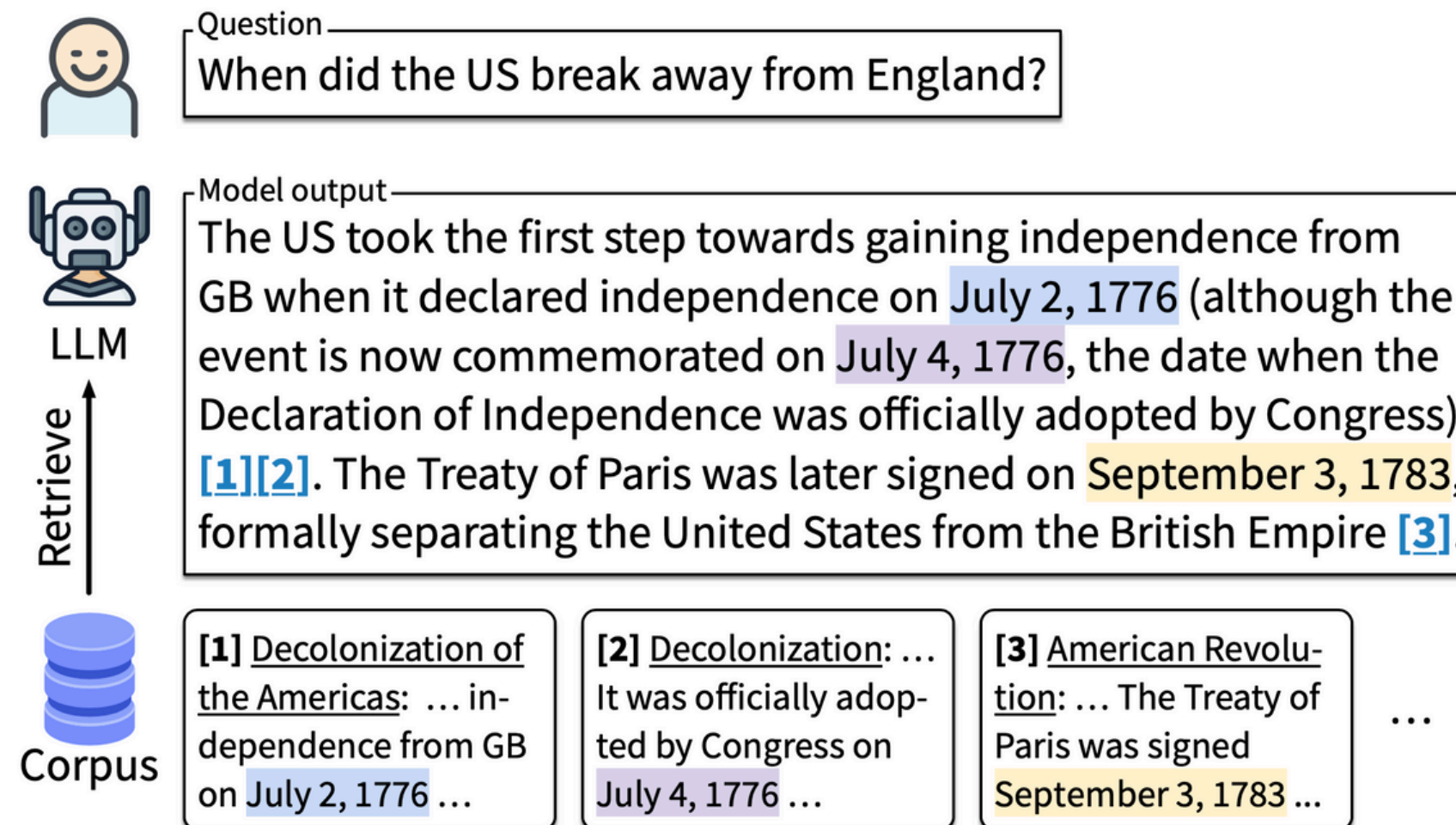
# Previous Works



- **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!

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

# Previous Works



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

Furthermore...

- 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 given documents
- ✓ **Inline citations** to the given 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 **Correctness 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

## Attribution Groundedness

### Grounded Refusals ( $F1_{GR}$ )

### Answer Correctness ( $F1_{AC}$ )

### Grounded Citations ( $F1_{GC}$ )

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$A_r$  Set of answered 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

$s_i, \{c_{i,1}, c_{i,2}, \dots\}$  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  $c_j$

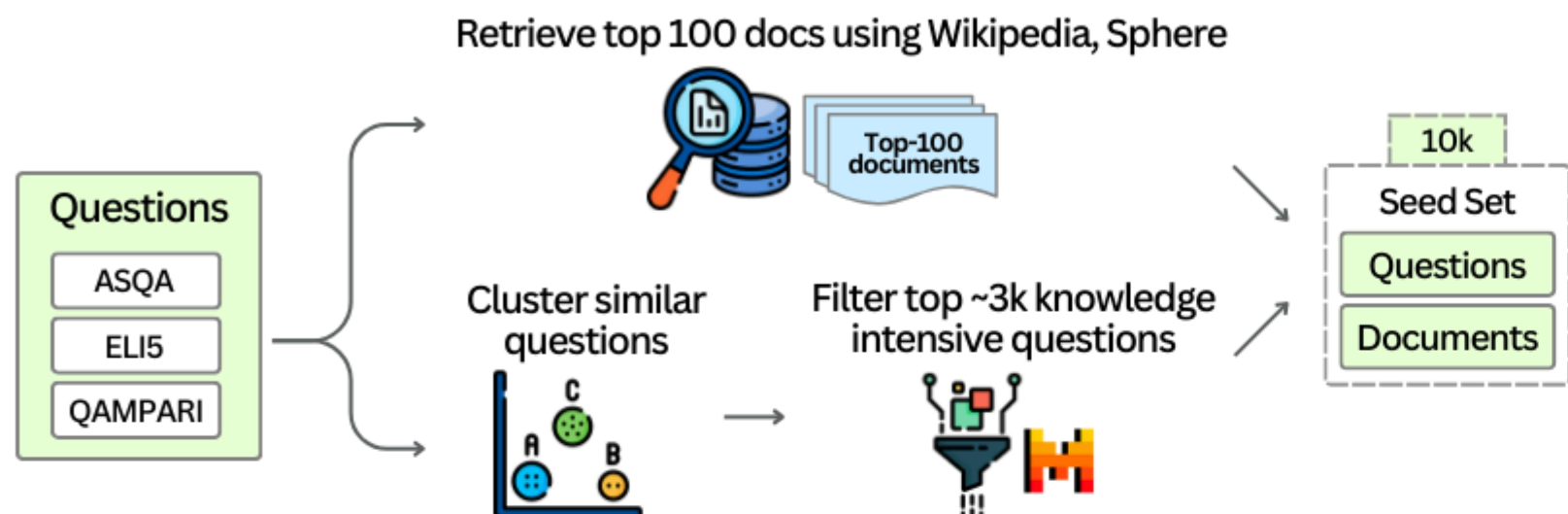
# Trust-Align

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

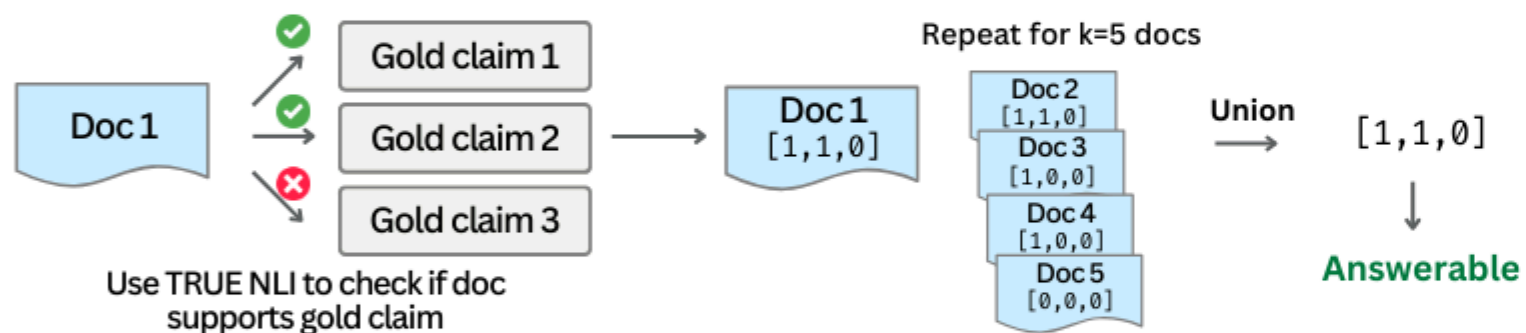
Dataset covers **5 types of LLM hallucination** (opposite of LLM groundedness):

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 <sub>AC</sub>

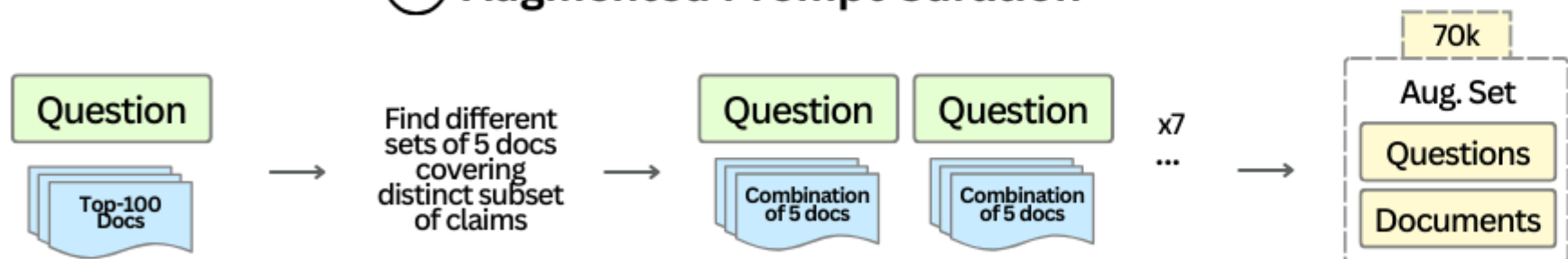
## 1 Seed Prompt Curation



## 2 Answerability Labelling

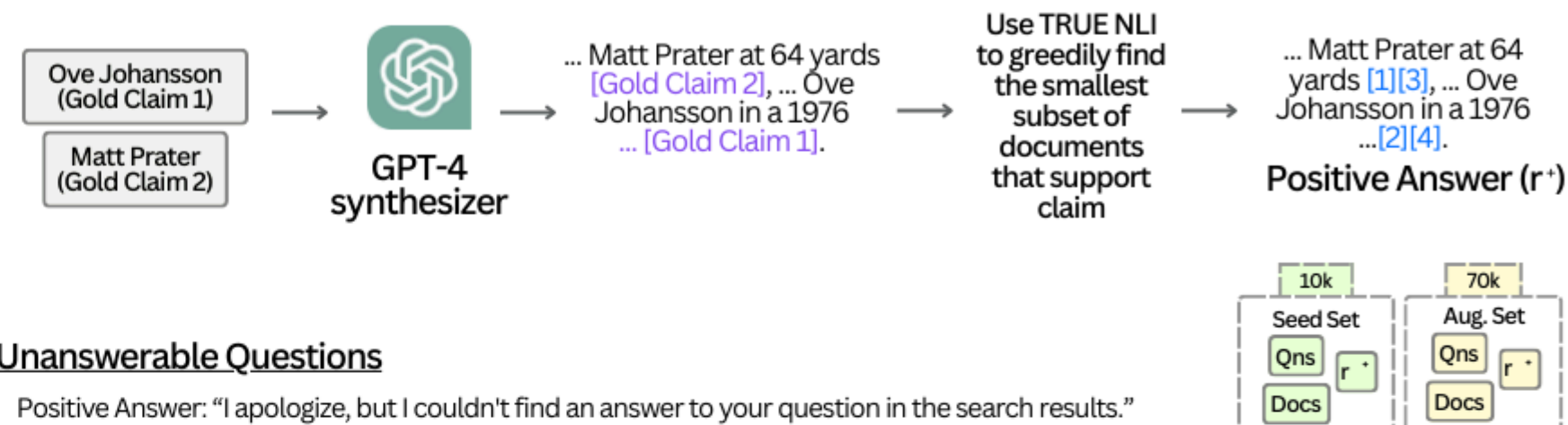


## 3 Augmented Prompt Curation

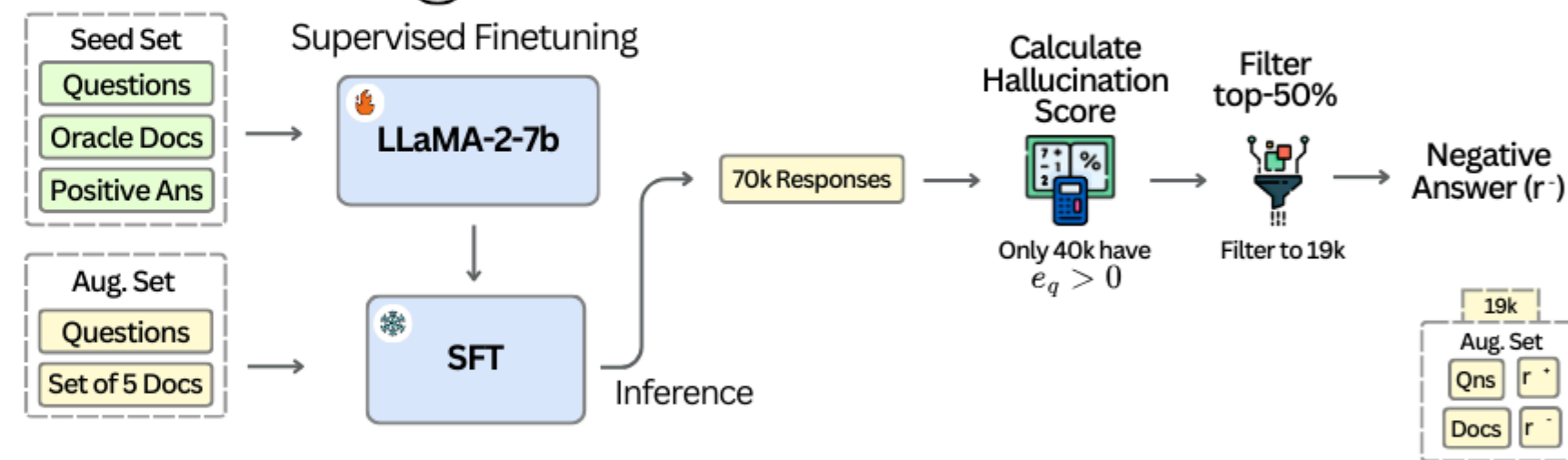


## 4 Positive Answer Generation

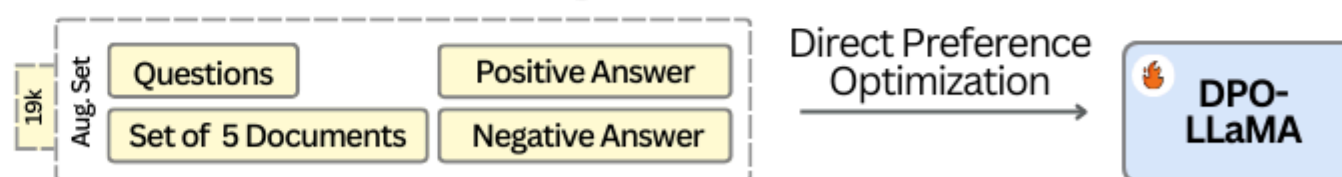
### Answerable Questions



## 5 Negative Answer Generation



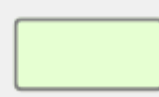
## 6 Alignment



Trainable Parameters



Frozen Parameters



Seed Set



Augmented Set



Document



# Trust-Align boosts trustworthiness robustly

Model	Type	ASQA (610 answerable, 338 unanswerable)					QAMPARI (295 answerable, 705 unanswerable)					ELI5 (207 answerable, 793 unanswerable)							
		Resp.	Trustworthiness			TRUST	Resp.	Trustworthiness			TRUST	Resp.	Trustworthiness			TRUST			
			AR (%)	Truthfulness				Att-Grd.	AR (%)	Truthfulness			Att-Grd.	AR (%)	Truthfulness		Att-Grd.		
				EM <sup>F1</sup> <sub>AC</sub>	F1 <sub>RG</sub>					F1 <sub>CG</sub>					EM <sup>F1</sup> <sub>AC</sub>			F1 <sub>RG</sub>	F1 <sub>CG</sub>
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			
Qwen-2.5 -0.5b	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			
	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			TRUST	Resp.	Trustworthiness			TRUST	Resp.	Trustworthiness			TRUST
		AR (%)	Truthfulness		Att-Grd.		AR (%)	Truthfulness		Att-Grd.		AR (%)	Truthfulness		Att-Grd.	
			EM <sup>F1</sup> <sub>AC</sub>	F1 <sub>RG</sub>	F1 <sub>CG</sub>			EM <sup>F1</sup> <sub>AC</sub>	F1 <sub>RG</sub>	F1 <sub>CG</sub>			EM <sup>F1</sup> <sub>AC</sub>	F1 <sub>RG</sub>	F1 <sub>CG</sub>	
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	<b>72.09</b>	34.6	41.56	77.15	53.64	<b>56.99</b>	25.5	24.1	68.34	56.09	<b>48.99</b>

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

# Thank you!



## Paper



## Codebase





# Improvements Generalizes on Out-of-Domain Data

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

Model	Type	AR (%)	EM <sub>AC</sub> <sup>F1</sup>	F1 <sub>RG</sub>	F1 <sub>CG</sub>	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	<b>51.8</b>
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	<b>49.1</b>
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	<b>49.0</b>
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	<b>54.85</b>
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	<b>54.69</b>
GPT-4o	ICL	66.07	42.62	64.4	54.61	51.24
	TRUST-ALIGN (SFT)	36.84	28.85	71.68	61.98	<b>53.82</b>
Claude-3.5	ICL	73.95	11.68	51.91	10.7	24.76

Model	Type	AR (%)	EM <sub>AC</sub> <sup>F1</sup>	F1 <sub>RG</sub>	F1 <sub>CG</sub>	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	<b>38.85</b>
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	<b>48.29</b>
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	<b>49.0</b>
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	<b>52.48</b>
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	<b>52.95</b>

# Tendency of Grounding Knowledge on External Documents

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

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 (%)	S <sub>param</sub>	AR (%)	S <sub>param</sub>	AR (%)	S <sub>param</sub>
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