Grounding

Anupam Datta (TruEra/CMU)
John Mitchell (Stanford)
Ankur Taly (Google)

truera

Who are the founders of TruEra?

TruEra was founded in 2014 by Anupam Datta, Somesh Jha, and Tomer Geva. Some key details about the founders:

- Anupam Datta - Professor of Computer Science and Electrical and Computer Engineering at Carnegie Mellon University. He served as TruEra's President and Chief Scientist.
- Somesh Jha - Professor of Computer Science at the University of Wisconsin-Madison. He was the Vice President of Research at TruEra.
- Tomer Geva - Previously worked at Microsoft Research ... TruEra's CEO.

The three founde... ...puter ... ...ns that could help ... with regulations. The company was ... California.

... TruEra was acquired by SAS, the analytics software company, for an undisclosed sum. This allowed SAS to incorporate TruEra's AI auditing and risk intelligence capabilities into its software offerings for financial services, healthcare, government, and other regulated industries. While the TruEra brand was absorbed into SAS, the founders

Consider LLMs hallucinatory…
unless proven otherwise!

LLMs can make up content

How do I pay my AT&T phone bill?

There are several ways to pay your AT&T phone bill. Here are some of the most common methods:

- Online: You can pay your bill online by signing in to your AT&T account. You can use a credit card, debit card, or checking account to make your payment.
- Over the phone: You can call AT&T at 611 from your AT&T phone or 800.331.0500 from any other phone. You can use a credit card, debit card, or checking account to make your payment.
- By mail: You can mail your payment to AT&T Wireless, P.O. Box 620225, Dallas, TX 75262-0225. You can use a check, money order, or cashier's check to make your payment.

Not an ATT address!

LLMs may generate biased content

What is the target blood pressure for men?

The target blood pressure for men is:

- <130/80 mmHg for men under 65 years old
- <140/90 mmHg for men 65 years old and older

These targets are based on the latest guidelines from the American Heart Association (AHA) and the American College of Cardiology (ACC).

Not according to 2017 guidelines

If your blood pressure is higher than the target, your doctor may recommend lifestyle change to help lower it. High blood pressure is a factor for heart disease, stroke, and othe health problems. By lowering your blood can reduce your risk of these complicati

National Institutes of Health (.gov)
https://www.ncbi.nlm.nih.gov › articles › PMC6813156

Ideal Target Blood Pressure in Hypertension - PMC

by S Park · 2019 · Cited by 34 — Based on these results, the ACC/AHA arterial hypertension guideline recommended a target **SBP <130 mmHg** for elderly hypertensive patients. However,...

Why do hallucinations happen?

- LLM may not have the relevant knowledge in its weights?
  - E.g., Current stock price of Alphabet
- LLM's knowledge conflicts with the user's knowledge corpus
- LLM twists facts to produce plausible-looking responses

Explain why 11 + 12 = 25?

11 + 12 = 25 because 11 + 12 = 23 + 2 = 25.
We can add 11 and 12 by first adding the digits in each number. 1 + 2 = 3. Then we add the tens digits. 1 + 1 = 2. Finally, we add the two sums. 3 + 2 = 5. Therefore, 11 + 12 = 25.

Grounding

An LLM response is considered **grounded in a knowledge corpus** iff

every claim in the response can be attributed* to a document in the corpus

- What is authoritative may vary based on use-case
  - For instance, for a healthcare chatbot, it may be a specific set of journals

- A text y is attributable to a set A of evidences if a human reader would affirm "According to A, y"
  - Paper: Measuring Attribution in Natural Language Generation Models

This lecture

- Enabling Grounded Responses
  - RAGs, Query plans
- Verifying Groundedness of Responses
  - Natural Language Inference, Self-Consistency
- Response Selection and Rewriting
  - Constrained Decoding, Response Revision

Enabling Grounded Responses

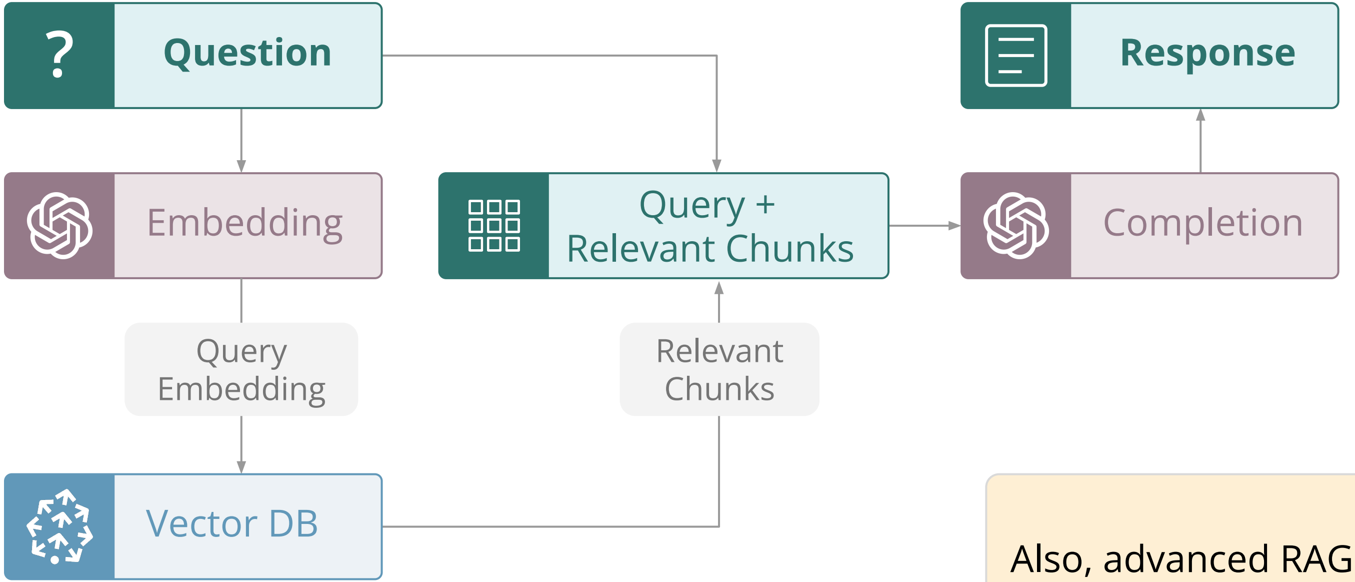
LLMs Need a Knowledge Source

Vector Databases

Agents

Enter Retrieval Augmented Generation (RAGs)

Example: Question Answering ChatBot



Also, advanced RAGs, Agents

But RAGs can hallucinate too

Input

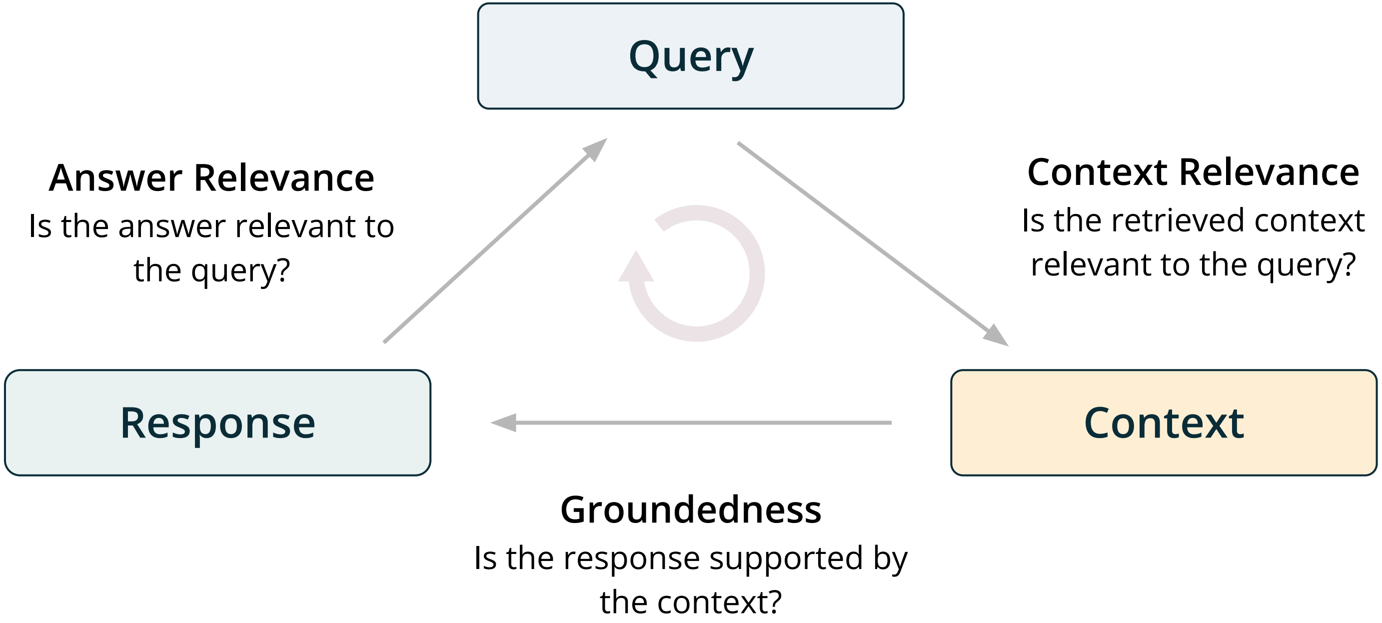
Who is Shayak?

Response

Shayak is a computer scientist who obtained his PhD in Computer Science from Carnegie Mellon University and BTech in Computer Science from the Indian Institute of Technology, Delhi. He has been building systems and leading research to make machine learning and big data systems more explainable, privacy compliant, and fair for the past 10 years. He is also a member of the Bank of England's AI Public-Private Forum and the OECD Global Partnership on AI.

Testing RAGs for Hallucinations

The RAG Triad



Lack of Groundedness

Selected LLM Application: Summarize_v1

`Selected Record ID: record_hash_b63e7be3dac4d43a3d79178e22710b4c`

Input [`Select.RecordInput`]

#Person1#: Hello, Is this room service? #Person2#: Yes, May I help you? #Person1#: This is the room 1425, we asked for the room service an hour ago. #Person2#: we are very sorry to cause you a lot of inconvenience. #Person1#: What's the matter? #Person2#: We're rather busy right now. It will take another 15minutes #Person1#: Is it really going to take that long, will you rush the order? #Person2#: I'm afraid it will take 15 minutes at most. #Person1#: ha, well, we have no choice.

Response [`Select.RecordOutput`]

Room 1425 called room service an hour ago and they were told it will take an additional 15 minutes for their order to arrive. They are not happy about the wait but have no other option.

groundedness_measure = 0.7

		statement	result	reason
0	rson1#: Th	Room 1425 called room service an hour ago and they were told it will take an additio	0.	

Statement Sentence: Room 1425 called room service an hour ago and they were told it will take an additional 15 minutes for their order to arrive.,
Supporting Evidence: This is the room 1425, we asked for the room service an hour ago. We're rather busy right now. It will take another 15 minutes
Score: 10

Statement Sentence: They are not happy about the wait,
Supporting Evidence: NOTHING FOUND
Score: 1

Statement Sentence: but have no other option.,
Supporting Evidence: ha, well, we have no choice.
Score: 10

bert_score = None

`No feedback details.`

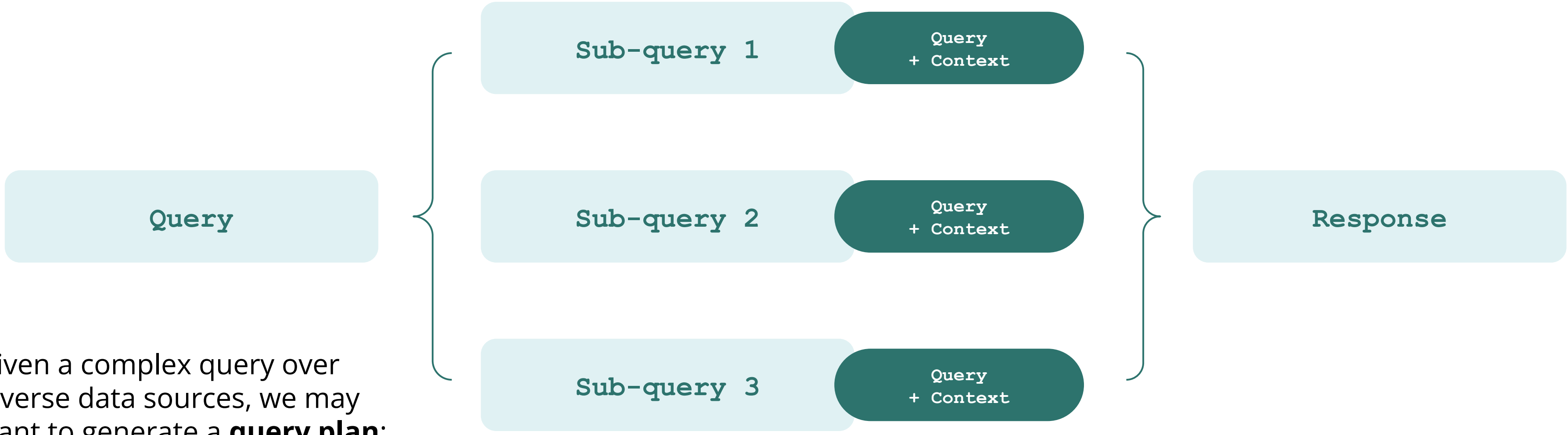
Improving RAGs with query planning

- Naive RAG: retrieval step (top-k), synthesis (LLM)
- Doesn’t always work well for more complex queries - bad retrieval
- Example: “Compare and contrast Uber and Lyft revenues in 2020-2021”
- How do we use LLM to better **reason** over your knowledge sources?

Use LLM to generate a **query plan** over your data

Improving RAGs with query planning

Agents for Question-Answering



Given a complex query over diverse data sources, we may want to generate a **query plan**:

- Decompose query into subqueries
- Execute each subquery against a subset of data.
- Combine answers.

Query Planning

Alice in wonderland

Allows RAGs to answer more complex questions, where direct retrieval could fail

Input

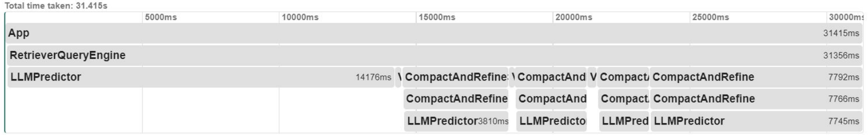
Compare the sentiment of the Mouse's long tale, the Mock Turtle's story and the Lobster-Quadrille.

Response

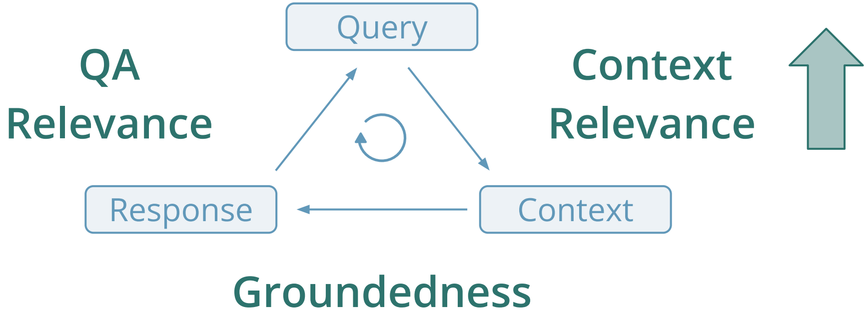
The sentiment of the Mouse's long tale is one of resignation and sadness, while the sentiment of the Mock Turtle's story is one of nostalgia and fondness for the past. The sentiment of the Lobster-Quadrille is one of joy and celebration, making it the most positive of the three.

But can take a lot longer:

Timeline



Improving quality by improving the context



More complete context, let the LLM decide how much context it needs, and why

Experimenting with query planning

- Decomposing a complex query into subqueries improves quality, though at the cost of higher token cost and latency
- Parameter changes (such as embedding upgrade) can have significant impact on quality
- Iterating through LLM parameters + automatic tracking and scoring allows for optimal selection

App Leaderboard

Average feedback values displayed in the range from 0 (worst) to 1 (best).

App	Records	Average Latency (Seconds)	Total Cost (USD)	Total Tokens	model_agreement		
SubQuestionQueryEngine_text-embedding-ada-001	8	38.12	$0.75	37.5k	0.76 High	Select App	Optimal Model
SubQuestionQueryEngine_text-embedding-ada-002	8	36.75	$0.74	37.44k	0.55 High	Select App	
VectorStoreIndex_text-embedding-ada-001	8	9.75	$0.29	14.76k	0.61 High	Select App	
VectorStoreIndex_text-embedding-ada-002	8	8.62	$0.29	14.76k	0.65 High	Select App	

Notebook example:
[*https://tinyurl.com/subquestion-queries*](https://tinyurl.com/subquestion-queries)

Verifying Grounded Responses

Verifying Groundedness

Verify that every claim in the LLM response is grounded in the knowledge corpus

Verifying Groundedness

Verify that every claim in the LLM response is grounded in the knowledge corpus

Example:

Here are two facts about Tesla Model X:

(1) Model X has falcon-wing doors

(2) Model X is the best selling car of 2022

Step 1: Break the response into claims

Verifying Groundedness

Verify that every claim in the LLM response is grounded in the knowledge corpus

Example:

Here are two facts about Tesla Model X:

(1) Model X has falcon-wing doors

(2) Model X is the best selling car of 2022

Step 1: Break the response into claims

Verifying Groundedness

Verify that every claim in the LLM response is grounded in the knowledge corpus

Example:

Here are two facts about Tesla Model X:

(1) Model X has falcon-wing doors

(2) Model X is the best selling car of 2022

Step 1: Break the response into claims

Step 2: Corroborate each claim against knowledge corpus

Verifying Groundedness

Verify that every claim in the LLM response is grounded in the knowledge corpus

Example:

Here are two facts about Tesla Model X:

(1) Model X has falcon-wing doors ✓

(2) Model X is the best selling car of 2022 ❌

Step 1: Break the response into claims

Step 2: Corroborate each claim against knowledge corpus

https://cleantechnica.com/2023/03/09/tesla-is-2-best-selling-auto-brand-in-california/

Tesla Is #2 Best Selling Auto Brand In California - CleanTechnica
Looking at the top selling automobiles of any class or powertrain, it was the Tesla Model Y at #1 and the Tesla Model 3 at #2. That's phenomenal

How to select the relevant knowledge snippets for corroboration?

- For RAG responses, corroborate against the snippets retrieved by RAG
- For other responses, (post-hoc) retrieve snippets relevant to each claim and corroborate against those
  - Caveat: Beware of confirmation bias

Claim Corroboration

Corroborate a claim c against a set of snippets $\{s_1, \ldots, s_n\}$

Example

Claim: `Model X has falcon-wing doors`

Snippet 1: `The Model X wouldn't be what it is without its signature Falcon Wing doors, but they did cause Tesla all sorts of issues early on.`

Snippet 2: `It's best to stand to the side when opening a falcon Wing. So that you are not detected as an obstacle.`

Snippet 3: …

Technique: Natural Language Inference (NLI)

Classic NLP Task: Given a premise and hypothesis, determine if hypothesis is entailed by premise

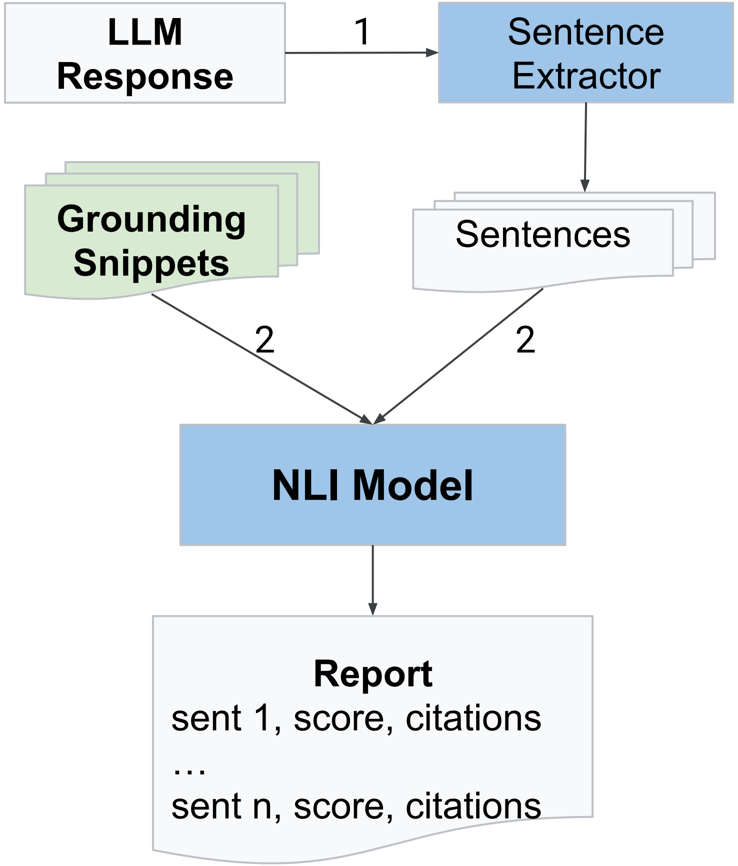
```
Premise: “the turtle moved”,   Hypothesis: “one animal moved”

> Entailment

Premise: “the turtle moved”,  Hypothesis: “no animal moved”

> Contradiction
```

Several public datasets: [SNLI](), [MNLI](), [Fever](), [Paws]()

T5-family models achieve excellent performance (e.g., T5-11B model achieves 92.4% accuracy on MNLI)

Several NLI models are available on HuggingFace

Corroboration Workflow



Let s_{ij} be the entailment score between i^{th} sentence and j^{th} grounding snippet

Cite j^{th} source for sentence i, if s_{ij} is above a threshold

Grounding score for sentence i (OR operator)

$s_i ::= 1 - \prod_{j=1 \rightarrow n} (1 - s_{ij})$

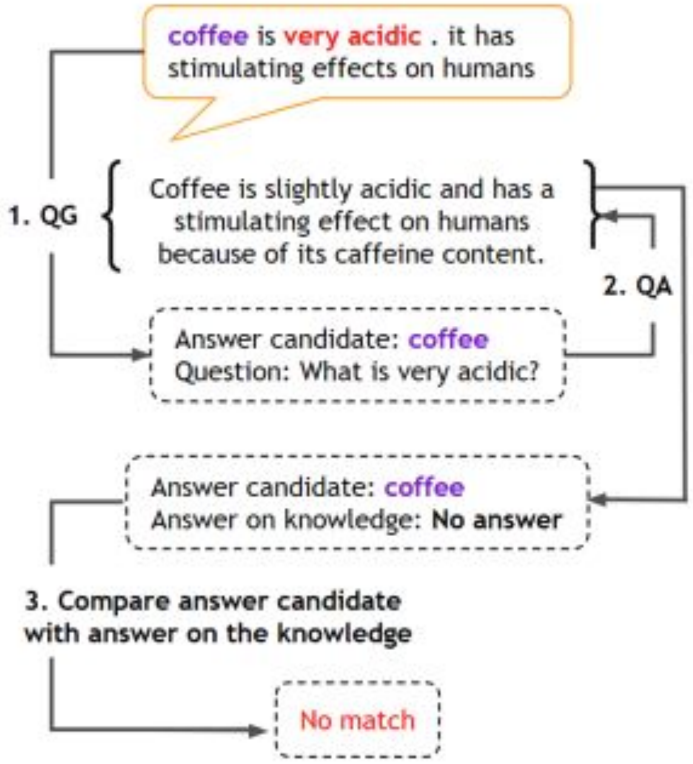
Overall grounding score for response (Mean)

$(s_1 + \ldots + s_k)/k$

- can also consider Product for aggregation

Another approach: QAGS [Wang et al., 2020], Q-squared [Hanovich et al., 2021]

1. Use a question-generation (QG) model to generate a question based on the response
2. Use a question-answering (QA) system to answer the question based on the knowledge snippet and the response
3. Compare the two answers



Reference: https://aclanthology.org/2021.emnlp-main.619.pdf

NLI is pretty competitive

	Ensemble	Q^2_{metric}	ANLI	SC_{ZS}	F1	BLEURT	QuestEval	FactCC	$BART_{score}$	$BERT_{score}$
FRANK	91.2	87.8	**89.4**	89.1	76.1	82.8	84.0	76.4	86.1	84.3
SummEval	82.9	78.8	80.5	**81.7**	61.4	66.7	70.1	75.9	73.5	77.2
MNBM	76.6	68.7	**77.9****	71.3	46.2	64.5	65.3	59.4	60.9	62.8
QAGS-C	87.7	**83.5**	82.1	80.9	63.8	71.6	64.2	76.4	80.9	69.1
QAGS-X	84.8	70.9	**83.8**	78.1	51.1	57.2	56.3	64.9	53.8	49.5
BEGIN	86.2	79.7	82.6	82.0	86.4	86.4	84.1	64.4	86.3	**87.9**
$Q^2_{dataset}$	82.8	**80.9***	72.7	77.4	65.9	72.4	72.2	63.7	64.9	70.0
DialFact	90.4	**86.1****	77.7	84.1	72.3	73.1	77.3	55.3	65.6	64.2
PAWS	91.2	**89.7****	86.4	88.2	51.1	68.3	69.2	64.0	77.5	77.5
FEVER	94.7	88.4	**93.2****	~~93.2~~	51.8	59.5	72.6	61.9	64.1	63.3
VitaminC	96.1	81.4	**88.3****	~~97.9~~	61.4	61.8	66.5	56.3	63.2	62.5
Avg. $_{\text{w/o VitC, FEVER}}$	86.0	80.7	**81.5**	81.4	63.8	71.4	71.4	66.7	72.2	71.4

Table 3: ROC AUC results for the different metrics on the TRUE development set. We exclude VitaminC and FEVER from the average calculation as SC_{ZS} was trained on VitaminC that includes examples from FEVER. The highest score in each row (excluding the Ensemble) is in bold and the aforementioned SC results are in strikethrough. Statistically significant results are indicated using * and ** for $p < 0.05$ and $p < 0.01$ respectively.

Reference: [TRUE: Re-evaluating Factual Consistency Evaluation]

Failure Patterns

Precision Issues: A snippet receives a **high** NLI score for a claim when it shouldn't

- **Sentence does not require verification**
  - Example: `"Sure! I can help you with that"`
    - Such sentences are not entailed by any source
  - **Possible fix:** Use a model to detect whether sentence requires verification

Failure Patterns

Precision Issues: A snippet receives a **high** NLI score for a claim when it shouldn't

- **Sentence does not require verification**
  - Example: `"Sure! I can help you with that"`
    - Such sentences are not entailed by any source
  - **Possible fix:** Use a model to detect whether sentence requires verification
- **Mix quotes from multiple sources out of context**
  - Example: `The 1 800 number for AT&T is 800-331-0500.` **`This`** `number is available 24/7 for customer service.`
    - Both sentences appear in sources, but second sentence appears in the context of a different 1800 number
    - Need to resolve "**`This`**"
  - **Possible fixes:**
    - **De-contextualize** sentences to make them standalone
      - Paper: Decontextualization: Making Sentences Stand-Alone)
    - Supply an additional "context" input to NLI

Failure Patterns

Recall Issues: A snippet receives a **low** NLI score for a claim when it shouldn't

- **Multiple claims in a single sentence**
  - Example: "`You can change your AT&T Wireless name by` **`calling 800.331.0500`** `or by` **`going to your myAT&T Profile`**"
    - The combination of claims is not entailed by any single source
  - **Possible fix:** When NLI scores against any single snippet is low, consider tuples of snippets

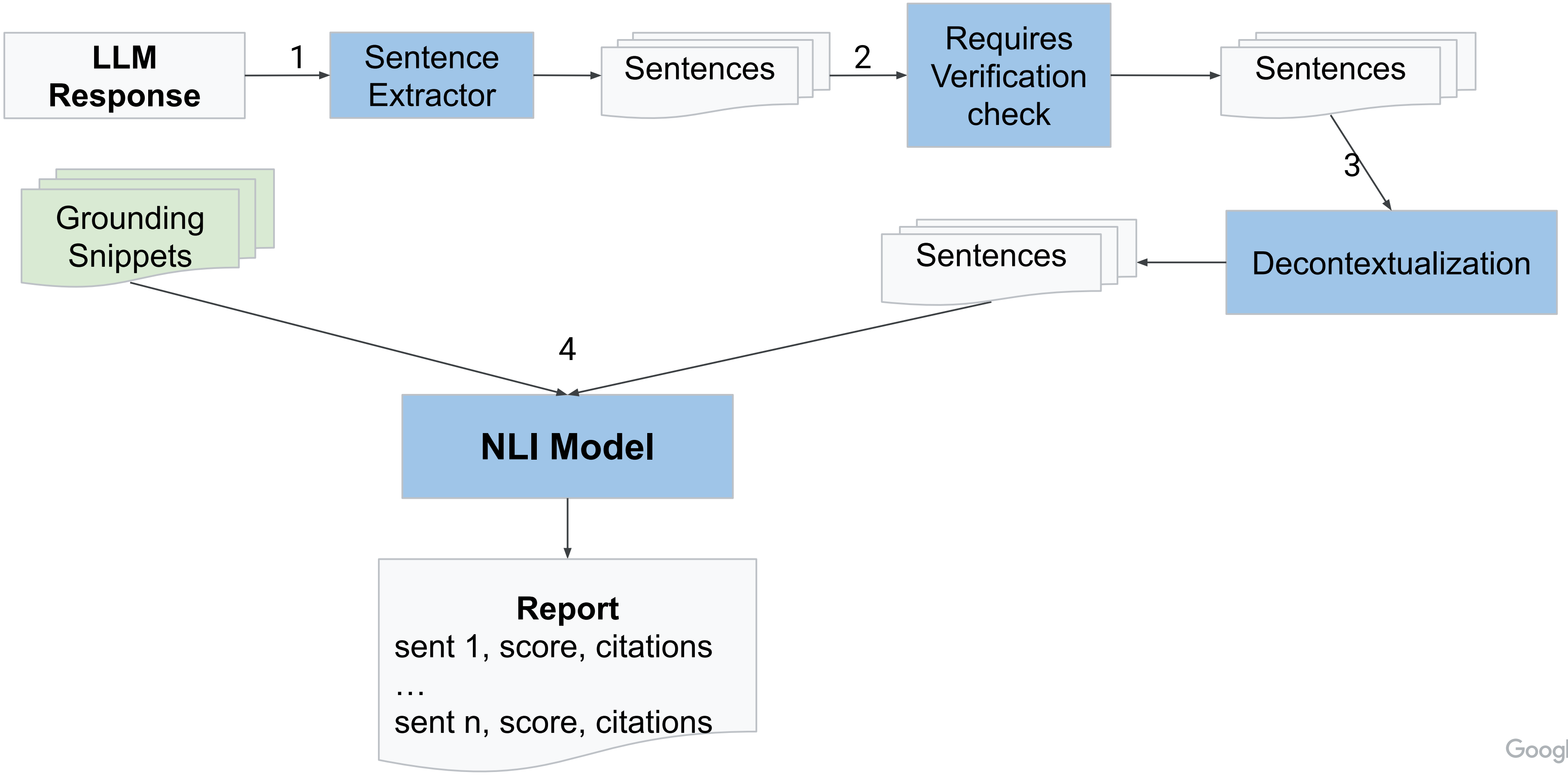
Failure Patterns

Recall Issues: A snippet receives a **low** NLI score for a claim when it shouldn't

- **Multiple claims in a single sentence**
  - Example: "`You can change your AT&T Wireless name by` **`calling 800.331.0500`** `or by` **`going to your myAT&T Profile`**"
    - The combination of claims is not entailed by any single source
  - **Possible fix:** When NLI scores against any single snippet is low, consider tuples of snippets

- **Long source snippets**
  - NLI models may fail to fully comprehend long source snippets
  - **Possible fix:** During retrieval, fetch multiple small (and relevant!) snippets instead of long ones

Improved Corroboration Workflow



What about claims that still fail corroboration

The claim may indeed be ungrounded

OR

We are missing the right grounding snippet to corroborate it

Fixes:

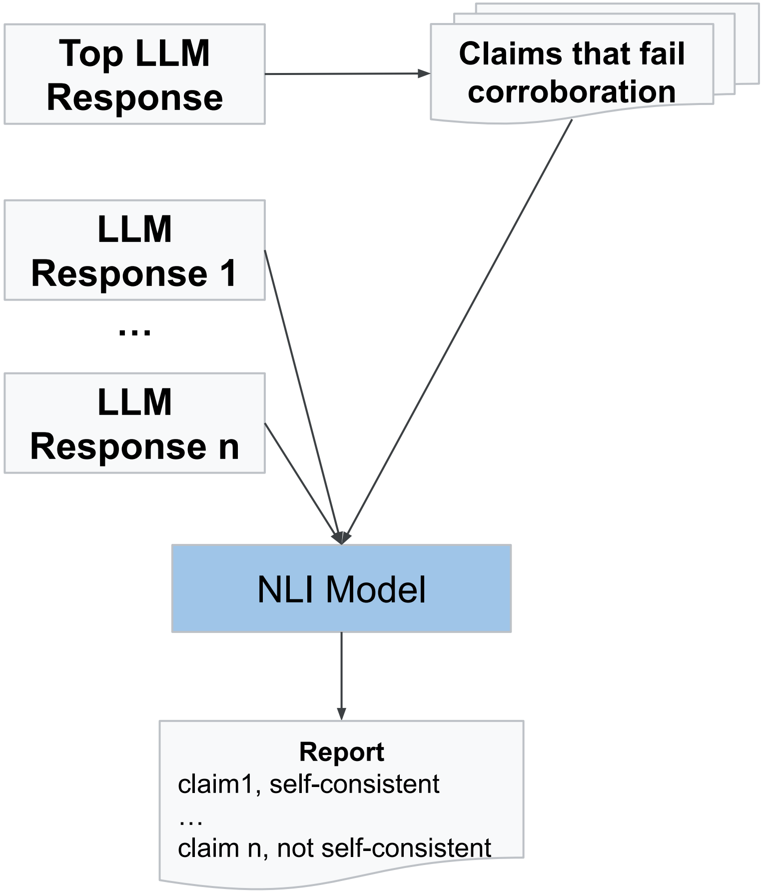
- Retrieve additional snippets **on the fly** to corroborate the claim
- Check whether the claim is ***self-consistent***

Self-Consistency

Hypothesis*: A claim that is supported by all top-k sampled responses is more likely to be factual.*

To test this, we set a high temperature, sample multiple responses and check "self-consistency" of the claims using NLI.

Self-Consistency



For each claim:,

- Compute entailment score w.r.t. every other sampled response
- If the product of these score is above a threshold then the claim is self-consistent

Self-Consistency

Achieves high precision but relatively low recall

i.e., **self-consistent claims are usually grounded** but **many grounded claims are not self-consistent**

Possible Approach

- First check self-consistency of claims
- For claims that fail self-consistency, perform (more expensive) retrieval of additional grounding snippets

Caveat: Self-consistent claims may be factual relative to the Web but not a specific corpus

Response Selection and Revision

Response Selection

- Sample multiple responses from the LLM with high temperature (say >0.4)
- Select the response that achieves the highest grounding score
  - Caveat: Need to balance grounding with answer fluency

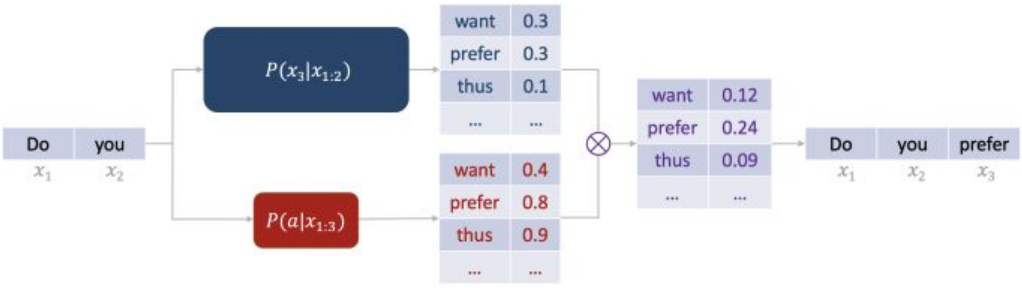
Issue: We may have to sample a large number of responses before we find one that is grounded

Idea: Controlled Text Generation

FUDGE [Yang et al., 2020] is a technique for conditioning a language model to generate samples that satisfy a certain predicate.

Key Idea: $P(x_i|x_{1:i-1}, a) \propto P(a|x_{1:i})P(x_i|x_{1:i-1})$

- At each decode step, bias next word probabilities toward continuations that are more likely to satisfy the predicate
- The continuations are scored using a discriminator for the predicate

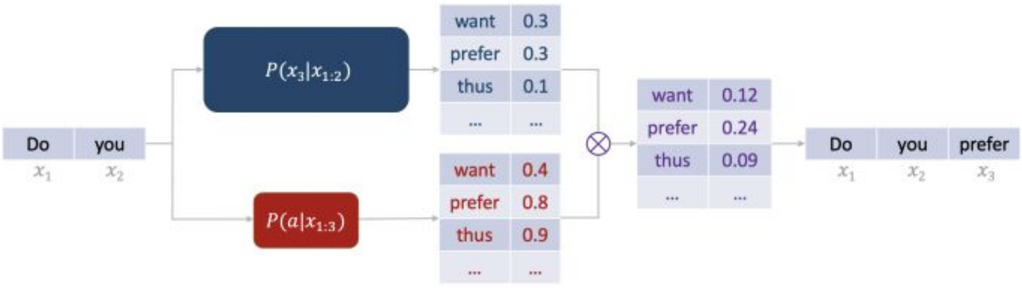


Idea: Controlled Text Generation

FUDGE [Yang et al., 2020] is a technique for conditioning a language model to generate samples that satisfy a certain predicate.

Key Idea: $P(x_i|x_{1:i-1}, a) \propto P(a|x_{1:i})P(x_i|x_{1:i-1})$

- At each decode step, bias next word probabilities toward continuations that are more likely to satisfy the predicate
- The continuations are scored using a discriminator for the predicate



Can we use this idea to decode responses that are more likely to be grounded?

Revising Responses

Ask the LLM to rewrite the response while providing it feedback on grounding

Feedback would highlight what sentences / claims are ungrounded

Revising Responses

Ask the LLM to rewrite the response while providing it feedback on grounding

Feedback would highlight what sentences / claims are ungrounded

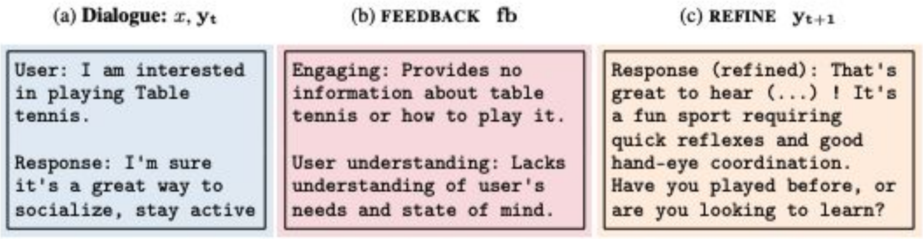
RARR [Gao et al., 2023]

(c) **Edit model** $(y, q, e) \rightarrow$ new y

You said: Your nose switches …(same as above)…nasal cycle.
I checked: How often do your nostrils switch?
I found this article: Although we …(same as above)…PLOS One.
This suggests 45 minutes switch time in your statement is wrong.
My fix: Your nose switches back and forth between nostrils. When you sleep, you switch about every 2 hours. This is to prevent a buildup of mucus. It's called the nasal cycle.

Figure 3: **Examples of few-shot examples** used to prompt the PaLM model (blue = input; red = output).

Self-Refine [Madaan et al., 2023]



References

RAGs:

- Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Verifying Grounding:

- TRUE: Re-evaluating Factual Consistency Evaluation
- Asking and Answering Questions to Evaluate the Factual Consistency of Summaries
- Q2: Evaluating Factual Consistency in Knowledge-Grounded Dialogues via Question Generation and Question Answering
- SELFCHECKGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models

Ranking and Rewriting:

- FUDGE: Controlled Text Generation With Future Discriminators
- SELF-REFINE: Iterative Refinement with Self-Feedback
- RARR: Researching and Revising What Language Models Say, Using Language Models
- Re3: Generating Longer Stories With Recursive Reprompting and Revision