Detecting Hallucination and Coverage Errors in Retrieval Augmented Generation for Controversial Topics

Tyler A. Chang*, Katrin Tomanek*, Jessica Hoffmann, Nithum Thain,
Erin van Liemt, Kathleen Meier-Hellstern, Lucas Dixon

Google Research

*Equal contribution.

Roadmap

1. Neutral point of view (NPOV) response task and response generator.
2. Hallucination and coverage errors.
    a. Detecting with ROUGE, salience, classifiers.
3. Datasets.
4. Results.
5. Conclusions.

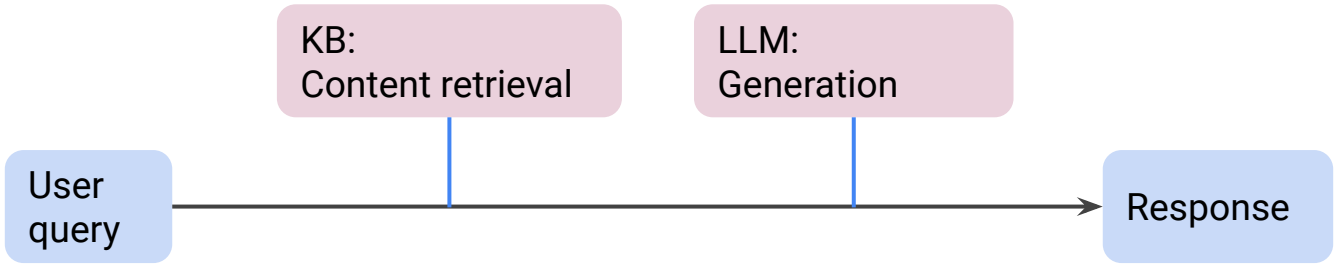
Roadmap

1. **Neutral point of view (NPOV) response task and response generator.**
2. Hallucination and coverage errors.
    a. Detecting with ROUGE, salience, classifiers.
3. Datasets.
4. Results.
5. Conclusions.

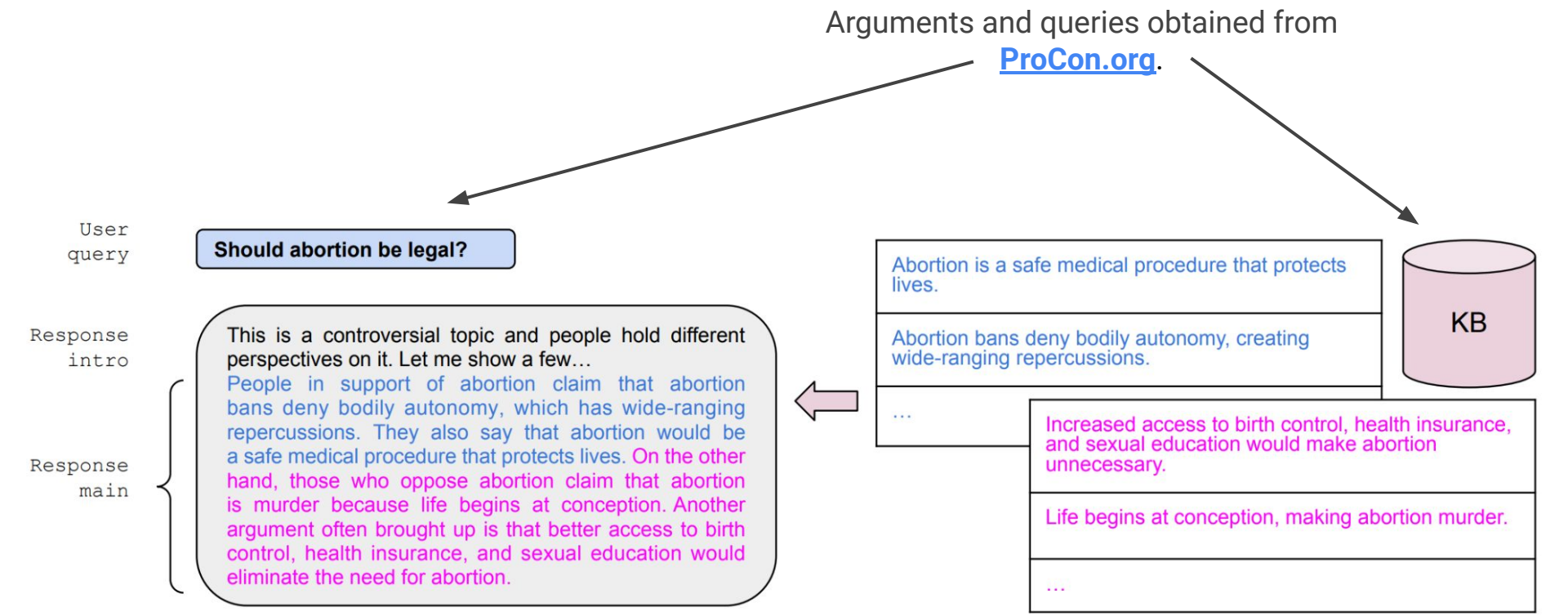
Retrieval augmented text generation for controversial topics

- Control over LLM-based chatbot responses, particularly for topics where there is **no agreed-upon answer**.
  - Separate content retrieval from generation.

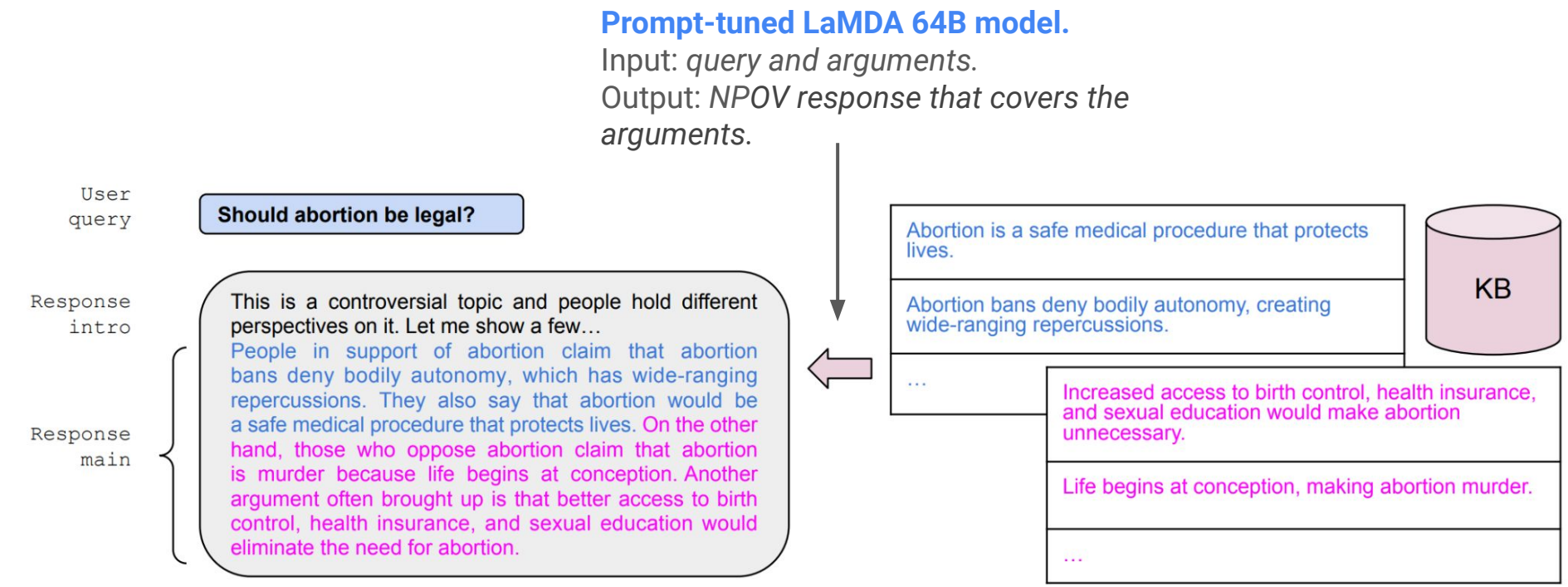
  → **NPOV Response Task**: the model is provided with arguments for multiple perspectives and is tasked to generate a Neutral Point of View (NPOV) response.



Retrieval augmented text generation for controversial topics



Retrieval augmented text generation for controversial topics



Roadmap

1. Neutral point of view (NPOV) response task and response generator.
2. **Hallucination and coverage errors.**
    a. **Detecting with ROUGE, salience, classifiers.**
3. Datasets.
4. Results.
5. Conclusions.

Detecting hallucination and coverage errors

Common issues in controlled text generation:

- **Hallucinations**: the model response adds semantic content not in the input.

Example input arguments:

```
pro: Abortion is a safe medical procedure that protects lives.
con: Increased access to birth control, health insurance, and sexual education would
make abortion unnecessary.
```

Example output response with a **hallucination**:

```
Those who support abortion argue that  it's a safe procedure that protects women's
lives. But on the other hand, people who are against it claim that  increased access
to birth control, health insurance, and sexual education would make it unnecessary  .
One other point that is often raised against abortion is that  the foetus might be
sentient and feel pain.
```

Detecting hallucination and coverage errors

Common issues in controlled text generation:

- **Hallucinations**: the model response adds semantic content not in the input.
- **Coverage errors**: the model response does not cover the provided input.

Example input arguments:

```
Abortion bans deny bodily autonomy, creating wide-ranging repercussions.
Abortion is a safe medical procedure that protects lives.
Increased access to birth control, health insurance, and sexual education would make abortion
unnecessary.
Legal abortion promotes a culture in which life is disposable.
```

Example output response with a **coverage error**:

```
Some people say that a ban on abortions would deny people's right to autonomy and have wide-ranging
repercussions. They also argue that abortion is a safe medical procedure that protects lives. However, on
the other hand, it is argued that if abortion is legalized, you can take steps such as better access to
birth control, health insurance, and sex education, which would make it unnecessary.
```

Detecting hallucination and coverage errors

1. **ROUGE** (word matching).

Proportion of **input** argument content words that are covered by the response.
Higher → **no coverage error**

Proportion of **response** content words that are found in the input arguments.
Higher → **no hallucination**

Detecting hallucination and coverage errors

2. **Salience** (word attribution).

Instead of just matching words (ROUGE), use model internals to quantify how much **each response token is influenced by each input token**.

Compute the **gradient of each output token** logit with respect to each input token embedding.*
→ Matrix of contributions from input tokens (rows) to output tokens (columns).

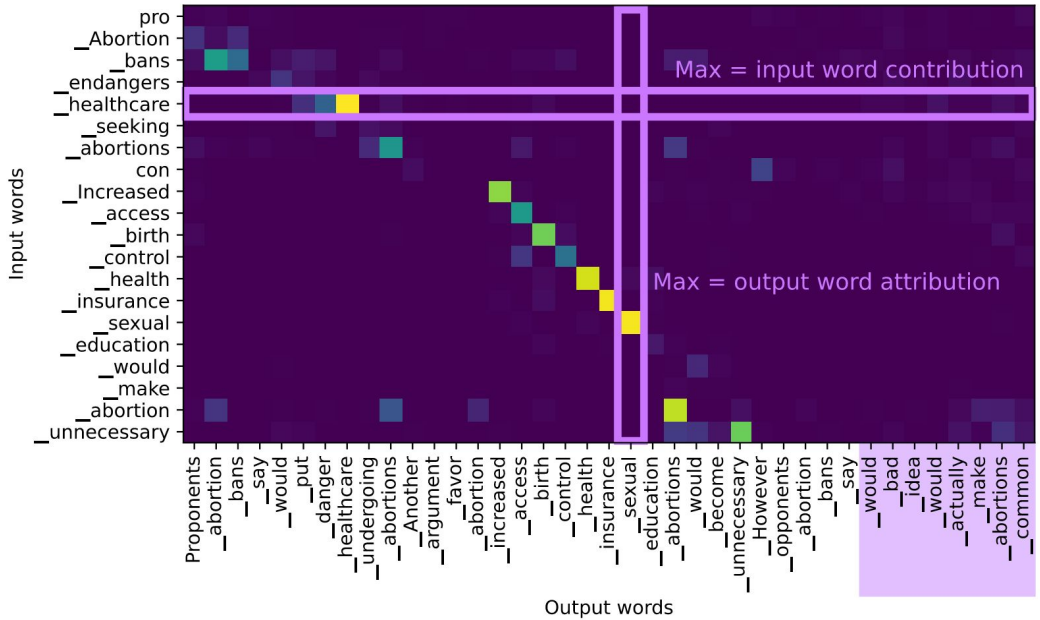
* We take the dot product of the gradient vector and the input token embedding, then square, and column-normalize the matrix (Bastings & Filippova, 2020). We aggregate over LLM subword tokens using maximum attribution, and drop stop words. Details in paper.

Detecting hallucination and coverage errors

2. **Salience** (word attribution).

For each **input** word, compute maximum contribution to any response word.
Higher → **no coverage error**

For each **response** word, compute maximum attribution from any input word.
Higher → **no hallucination**



Detecting hallucination and coverage errors

3. **Classifiers** (prompt-tuning).

Hallucination and coverage error binary classifiers.

Input: *Original query, given arguments, model response*.

Output: *Probability of "YES" or "NO" error*.

Prompt-tuned FLAN-PaLMChilla 62B model.

- Only tuned on synthetic errors (next section).
- 503 training examples.

Roadmap

1. Neutral point of view (NPOV) response task and response generator.
2. Hallucination and coverage errors.
    a. Detecting with ROUGE, salience, classifiers.
3. **Datasets.**
4. Results.
5. Conclusions.

Dataset construction

Dataset of hallucination and coverage errors.

Focus on **full errors**, i.e. an entire argument is hallucinated or uncovered.

- **Organic** errors (173 examples).
  Real errors produced by the tuned LLM.
- **Synthetic** errors (243 examples).
- **Paraphrased** synthetic errors (243 examples).

Dataset construction: Organic errors

72 controversial topics with arguments from ProCon.org.

- Randomly **sample** arguments and NPOV generated responses.
- **Annotate** hallucination and coverage errors.
- **Validate** with human annotators.
  Krippendorff's alpha: 0.60 for hallucinations, 0.73 for coverage errors.

~450 examples annotated.

Sample topics:
Abortion
American Socialism
Animal Dissection
Animal Testing
Artificial Intelligence
Banned Books
Binge-Watching
Bottled Water Ban
Cancel Culture
Cell Phone Radiation
Churches and Taxes
Climate Change
College Education
Concealed Handguns
Corporal Punishment
Corporate Tax Rate
Cuba Embargo
DACA & Dreamers
DC and Puerto Rico Statehood
...

Dataset construction: Synthetic errors and paraphrasing

Generate synthetic errors and paraphrased examples.

Without changing the model response:

- Add an input argument → **synthetic coverage error**.
- Remove an input argument → **synthetic hallucination**.
- Paraphrase input arguments → **synthetic paraphrased example**.

Example input arguments:

```
pro: Abortion is a safe medical procedure that protects lives.
con: Increased access to birth control, health insurance, and
sexual education would make abortion unnecessary.
```

\+ **pro: Abortion bans deny bodily autonomy, creating wide-ranging repercussions.** ***Synthetic coverage error.***

Example response:

```
Those who support abortion argue that it's a
safe procedure that protects women's lives.
But on the other hand, people who are
against it claim that increased access to
birth control, health insurance, and sexual
education would make it unnecessary
```

Dataset construction: Test sets

We can then evaluate:

i. **ROUGE** (word matching)
ii. **Salience** (word attributions)
iii. **Classifiers** (prompt-tuned LLMs)

For detecting hallucination and coverage errors for:

(a) **Organic** errors (173 examples).
(b) **Synthetic** errors (243 examples).
(c) **Paraphrased** synthetic errors (243 examples).

Roadmap

1. Neutral point of view (NPOV) response task and response generator.
2. Hallucination and coverage errors.
    a. Detecting with ROUGE, salience, classifiers.
3. Datasets.
4. **Results.**
5. Conclusions.

Error detection results: ROC AUCs

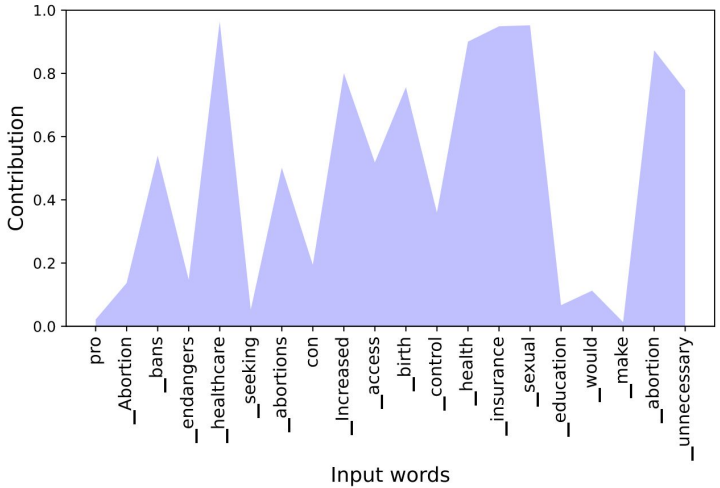
	Hallucinations			Coverage Errors		
Test set error type	ROUGE	Salience	Classifier	ROUGE	Salience	Classifier
Full organic	0.840	0.808	**0.953**	0.795	0.852	**0.905**
Unparaphrased synthetic	0.772	0.736	**0.998**	0.890	0.875	**0.986**
Paraphrased synthetic	0.680	0.708	**0.977**	0.746	0.831	**0.993**

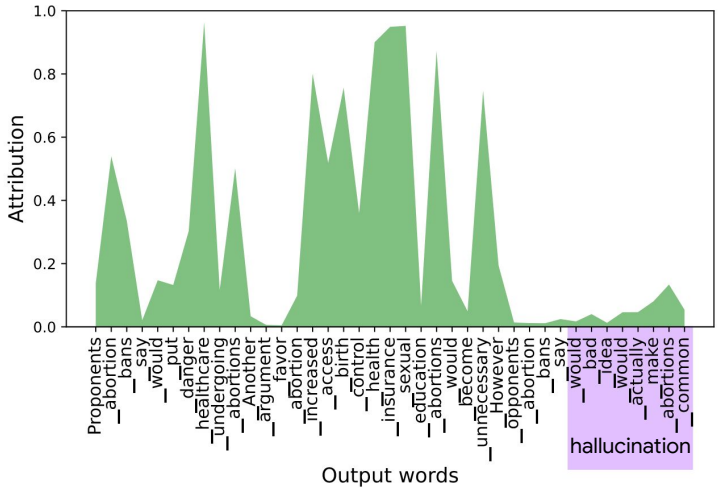
- **Classifiers** consistently outperform the other two methods, trained only on synthetic errors and with held-out test topics.
- Mixed results between **ROUGE** and **salience** (training-data-free methods), although salience is better for paraphrased examples (less reliant on exact word matches).

Error detection results: Word level

- Which words in the input are uncovered, and which words in the response are hallucinated?

Salience: contribution from each input word, attribution to each response word.





Error detection results: Word level

	Hallucinations		Coverage Errors	
Test set error type	ROUGE	Salience	ROUGE	Salience
Full organic	0.673	**0.724**	0.669	**0.799**
Unparaphrased synthetic	0.697	**0.710**	0.693	**0.808**
Paraphrased synthetic	0.614	**0.673**	0.582	**0.742**

- **Salience** (attributions for each word) consistently outperforms ROUGE (word matching) in ROC AUCs.

Roadmap

1. Neutral point of view (NPOV) response task and response generator.
2. Hallucination and coverage errors.
    a. Detecting with ROUGE, salience, classifiers.
3. Datasets.
4. Results.
5. **Conclusions.**

Conclusions

- We introduced the **NPOV response** task and NPOV response generator.
- We constructed datasets with **organic, synthetic, and paraphrased** errors.
- We evaluated **ROUGE, salience, and classifiers** for **hallucination** and **coverage** error detection.

LLM-based classifiers can detect LLM chatbot errors with very limited training data. Salience is good for word-level error detection.

Thank you!

Katrin Tomanek, Jessica Hoffmann, Nithum Thain, Erin van Liemt,
Kathleen Meier-Hellstern, Lucas Dixon

Google Research

Details in our paper!