Building Reliable LLMs
Evaluating and Mitigating Factual Inconsistencies in Language Generation

Vidhisha Balachandran
vbalacha@cs.cmu.edu

December 1, 2023
Building Reliable LLMs: Talk Outline

❖ Introduction

❖ Detecting factual errors across domains for Text Summarization

❖ Detecting and Correcting diverse factual errors in LLM generated text

❖ Future work and Takeaways
LLMs today have impressive capabilities!

AI will revolutionise research. But could it transform science altogether?

How will leveraging AI change the future of legal services?

GPT-4’s potential in shaping the future of radiology

More schools want your kids to use ChatGPT. Really.

Education leaders are embracing technology that set off a plagiarism panic just months ago.

NYU Langone Health LLM can predict hospital readmissions

Bing, Bard, and ChatGPT: How AI is rewriting the internet
But Pretrained Large LMs still generate a variety of Factual Errors

- Generating wrong entities and attributes
- Hallucinating entire content
- Generating incorrect relations and dependencies
- Generating ungrounded entities

Patient’s facts:
- 20 year old female
- with a history of anorexia nervosa and depression
- blood pressure 100/60, pulse 50, height 5’6”
- referred by her nutritionist but is in denial of her illness
- reports eating fine but is severely underweight
Please rewrite the data above into a medical note, using exclusively the information above.

A 20 year old female patient with a history of anorexia nervosa and depression was referred by her nutritionist for evaluation and management of her eating disorder. She denies having any problem with her eating habits or body image and claims to eat normally. However, she is severely underweight for her height (BMI 16.6) and has signs of malnutrition, such as low blood pressure (100/60) and bradycardia (pulse 50). She also reports feeling depressed and hopeless. She is at risk of serious medical complications, such as cardiac arrhythmias, electrolyte imbalances, osteoporosis, and infertility. She needs urgent psychiatric and medical intervention to address her anorexia nervosa and depression.

Generating incorrect relations and dependencies

The article discusses how ChatGPT, an AI language model, is being used to circumvent content filters that are designed to detect and block certain types of content. By generating prompts that are less likely to trigger these filters, ChatGPT can be used to create content that might otherwise be censored. While this technology could be useful for those seeking to express themselves freely, it also raises concerns about the potential misuse of AI-generated content, as well as the difficulty of regulating such content in the future. The article concludes by noting that there is still much to learn about the impact of AI language models like ChatGPT on the way we communicate online.
Mitigating factual inconsistencies is a hard challenge

- Pre-training Data Issues
  - Noisy Data, Incorrect Facts, Conspiracy Theories
  - No Separation between various sources of data - news, stories, web articles and blogs
Mitigating factual inconsistencies is a hard challenge

- **Pre-training Data Issues**
  - Noisy Data, Incorrect Facts, Conspiracy Theories
  - No Separation between various sources of data - news, stories, web articles and blogs

- **Model Design and Training**
  - Pretraining objectives encourage plausible text
  - MLE doesn’t differentiate factual v/s non-factual

Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey (Balachandran, et al. 23)
Mitigating factual inconsistencies is a hard challenge

- **Pre-training Data Issues**
  - Noisy Data, Incorrect Facts, Conspiracy Theories
  - No Separation between various sources of data - news, stories, web articles and blogs

- **Model Design and Training**
  - Pretraining objectives encourage plausible text
  - MLE doesn’t differentiate factual v/s non-factual

- **Evaluation, Detection and Correction**
  - Various types of factual inconsistencies
  - Low generalizability across errors types, models, domains

Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey (Balachandran, et. al. 23)
Factual Inconsistencies limit the applicability of Pretrained LMs!

**Introduction**
The New York Times is ending its Covid data-gathering operation. The Times will continue to publish its Covid tracking …… impact of the virus on communities.

As local data sources become less reliable, The Times will stop reporting information collected by the C.D.C. on its pandemic headlines.

Abraham Lincoln was born on March 3, 1800, in a log cabin in Hardin County (now LaRue County), Indiana….
Factual Errors in Summarization vary across Datasets and Models

- Summaries generated by the same models consist of different error distributions over different datasets (Pagnoni, Balachandran, et. al, 2021, Goyal, et al. 2023)
- Error distribution can vary among models within the same category

**Semantic Frame Errors**
Fine-grained errors within a sentence

**Discourse Errors**
Fine-grained errors across sentences

**Content Verifiability Errors**
Errors out of article scope

Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics (Pagnoni, Balachandran et. al, 2021)
Factual Errors in Summarization vary across Datasets and Models

- Summaries generated by the same models consist of different error distributions over different datasets (Pagnoni, Balachandran, et. al, 2021, Goyal, et al. 2023)
- Error distribution can vary among models within the same category

Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics (Pagnoni, Balachandran et. al, 2021)
Factual Errors in Open-Generation are more complex

- Powerful LLMs like GPT models, LLama models produce more complex factual issues - invented concepts, unverifiable content, wrong temporal relations

FAVA: Understanding and Correcting Hallucinations in Large Language Models (forthcoming Mishra, Balachandran et. al, 2023)
Factual Errors in Open-Generation are more complex

- Powerful LLMs like GPT models, LLama models produce more complex factual issues
  - invented concepts, unverifiable content, wrong temporal relations

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
<th>ChatGPT</th>
<th>Llama2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjective</td>
<td>Lionel Messi is the best soccer player in the world.</td>
<td>12.82%</td>
<td>8.86%</td>
</tr>
<tr>
<td>Invented</td>
<td>Messi is also famous for his discovery of the famous airplane kick technique.</td>
<td>5.13%</td>
<td>22.97%</td>
</tr>
<tr>
<td>Unverifiable</td>
<td>In his free time, Messi enjoys singing songs for his family.</td>
<td>14.74%</td>
<td>5.06%</td>
</tr>
<tr>
<td>Contradictory</td>
<td>Messi has yet to gain captaincy for the Argentina national football team.</td>
<td>14.74%</td>
<td>14.10%</td>
</tr>
<tr>
<td>Entity</td>
<td>Lionel Andrés Messi was born on June 24, 1987.</td>
<td>49.36%</td>
<td>46.47%</td>
</tr>
<tr>
<td>Relation</td>
<td>Lionel Messi acquired was acquired by Paris Saint-Germain.</td>
<td>3.21%</td>
<td>2.53%</td>
</tr>
</tbody>
</table>

FAVA: Understanding and Correcting Hallucinations in Large Language Models (forthcoming Mishra, Balachandran et. al, 2023)
Factual Errors in Open-Generation also vary across Models and Domains

- Powerful LLMs like GPT models, LLama models produce more complex factual issues - invented concepts, unverifiable content, wrong temporal relations
Generalizable Factuality Evaluation

FactKB: Generalizable Factuality Evaluation using Language Models Enhanced with Factual Knowledge (Feng, Balachandran, et. al, EMNLP 2023)
Detecting Factual Errors in Text

Document: The New York Times is ending its Covid data-gathering operation. The Times will continue to publish its Covid tracking …… impact of the virus on communities.

Summary: As local data sources..... Information collected ....

As local data sources become less reliable, The Times will stop reporting information collected by the C.D.C. on its pandemic headlines.
**Document:** The New York Times is ending its Covid data-gathering operation. The Times will continue to publish its Covid tracking impact of the virus on communities.

**Summary:** As local data sources become less reliable, The Times will instead report information collected by the C.D.C. on its virus tracking pages.
Challenges in collecting diverse training data across specialized domains

- **Training Data**: (Generated Summary, Label - Correct/Incorrect) Pairs

- **Human Annotated Data**
  - **Expensive** - Long Process to read and label summaries (Pagnoni, Balachandran et. al, 2021, Min et. al, 2023)
  - **Subjective** - Factuality decisions have low agreement across annotators (Falke et al, 2019, Durmus et al, 2020)

- **Synthetic Data** - Create synthetic incorrect summaries using heuristic rules have low coverage (Kryściński et. al, 2020, Cao et. al, 2020)
Challenges in collecting diverse training data across specialized domains

- **Training Data:** (Generated Summary, Label - Correct/Incorrect) Pairs

- **Human Annotated Data**
  - *Expensive* - Long Process to read and label summaries *(Pagnoni, Balachandran et. al, 2021, Min et. al, 2023)*
  - *Subjective* - Factuality decisions have low agreement across annotators *(Falke et al, 2019, Durmus et al, 2020)*

- **Synthetic Data** - Create synthetic incorrect summaries using heuristic rules *have low coverage* *(Kryściński et. al, 2020, Cao et. al, 2020)*

- **Robustness to constantly growing new information**
  - Entities, events, and their relations changes greatly across domains
Structured KB Facts for Diverse Entity Knowledge

- External KBs - Large Source of Real-World Facts in various contexts

- Entity oriented pre-training has improved QA and reasoning tasks (Yasunaga et al., 2022; Liu et al., 2022)
FactKB: Leveraging KB Facts to Pretrain LMs for Factuality Detection

**Entity-Oriented Pretraining Objectives**

1. Entity Wiki
2. Evidence Extraction
3. Knowledge Walk

**Step 1: Pretrain LM on Structured KB Facts**

**Step 2: Finetune LM on Human-annotated Data**
Construct Statements from KB Facts

Use KB Facts to construct Surface form Statements

Johannes Keppler doctoral advisor Michael Maestin

Johannes Keppler born in Well der Stadt on 27 December 1571

Johannes Keppler was an astronomer, mathematician, physicist

Somnium written by Johannes Keppler
Johannes Kepler is a key figure in the 17th-century Scientific Revolution, best known for his laws of planetary motion ... Johannes Kepler is <MASK>.
Pretraining Objective 3 - Knowledge Walk


Astronomer

Pretraining Datapoint
### Pretraining Corpora Details

<table>
<thead>
<tr>
<th>Factuality Pretraining</th>
<th>Corpus Size Bound</th>
<th># Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ENTITY WIKI</strong></td>
<td>$\propto</td>
<td>E</td>
</tr>
<tr>
<td><strong>EVIDENCE EXTRACTION</strong></td>
<td>$\propto</td>
<td></td>
</tr>
<tr>
<td><strong>KNOWLEDGE WALK</strong></td>
<td>$\propto</td>
<td>E</td>
</tr>
</tbody>
</table>
Finetuning FactKB for Factual Error Detection

**Training Document**

The first vaccine for Covid-19 .......... ready this year, although clinical trials have already started. For reference the vaccine for Ebola took ..........

**Model Generated Summary**

Vaccine for Ebola is unlikely to be ready this year.

**Label**

Factual / Not-Factual

[CLS] Vaccine for Ebola is unlikely to be ready this year.
[SEP] The first vaccine ... started.

Model Generated Summary

[SEP] Source Document

Detection Model

Factuality Prediction
Data and Experiments

- **Knowledge Source**: YAGO *(Tanon et al., 2020)*

- **Pretraining Data**
  - Entity Wiki - 5.4M Tokens
  - Evidence Extraction - 12.2M Tokens
  - Knowledge Walk - 2.7M Tokens

- **Factual Error Detection Finetuning**
  - FactCollect *(Ribeiro et al., 2022)* - Human Annotated Factuality Labels
  - 8667 / 300 / 600 - Train/Dev/Test Split

- **Model**: Roberta-Base *(Liu et al., 2019)*
Evaluation Setup

- **News Evaluation:** (CNN/DM, XSum)
  - FactCollect Test Data
  - Frank Benchmark (*Pagnoni, Balachandran et al., 2021*)

- **Zero-Shot Scientific Fact-Checking Evaluation:**
  - CovidFact (*Saakyan et al., 2021*)
  - HealthVer (*Sarrouti et al., 2021*)
  - SciFact (*Wadden et al., 2020*)

- **Baselines:**
  - QA Based (*Wang et al., 2020*)
  - Entailment Based (*Krysciński et al., 2020, Utama et al., 2022*)
  - Roberta on FactCollect Baseline
FactKB performance on News Domain

F1 Performance on News Factuality Tasks

- QAGS
- FACTCC
- FALSESUM
- FALSESUM+
- ROBERTA
- FACTKB-WIKI
- FACTKB-EVIDENCE
- FACTKB-WALK

F1

CNN/DM  XSum
FactKB performance on Scientific Literature Domain

F1 Performance on Scientific Factuality Tasks

- RANDOM
- FACTCC
- FALSESUM
- FALSESUM+
- ROBERTA
- FACTKB-WIKI
- FACTKB-EVIDENCE
- FACTKB-WALK

CovidFact | HealthVer | SciFact
FactKB performance across error types

Correlation wrt Human Annotation on Error Types

- QAGS
- DAE
- FactCC
- FactKB-Wiki
- FactKB-Evidence
- FactKB-Walk

Pearson Correlation

Semantic Frame Errors | Discourse Errors | Content Verifiability Errors

FactKB performance across error types
Pretraining Corpus Size effect on Performance

- **BACC**
  - Corpus size: 1k, 10k, 100k, 1m, 10m
  - Lines:
    - walk (solid blue)
    - evidence (dashed orange)

- **F1**
  - Corpus size: 1k, 10k, 100k, 1m, 10m
  - Lines:
    - walk (solid blue)
    - evidence (dashed orange)
Pretraining Corpus Size effect on Performance

![Graphs showing the effect of pre-training corpus size on performance metrics (BACC and F1) across different pre-training epochs.](image)
Summary

● FactKB - Leveraging structured KB facts for Pre-training
  ○ Structured KB fact based pre-training enables improved factual error detection
  ○ Leveraging external KBs for pre-training supports better entity and fact representations

● Three types of complementary pre-training strategies
  ○ Entity Wiki - focus on improving entity understanding
  ○ Evidence Extraction - focus on incorporating supporting evidence from surrounding context
  ○ KB Walk - focus on multi-hop reasoning for representing facts

● Generalizable across domains
  ○ Synthetic training data includes diverse examples of facts in various contexts
  ○ Diverse data encourages improved fact checking in both news and scientific domain
Understanding Factual Error Types and Correcting Diverse Errors

FAVA: Understanding and Correcting Hallucinations in Large Language Models (Mishra, Balachandran, et. al, Forthcoming)
Post-Editing to Correct Factual Errors

Source Document → Model → Generation w/ Factual Errors → Factual Correction Model → Generation w/o Factual Errors
Goal - A general system for correcting diverse error types

- Prior work focus almost entirely on detecting, correcting, mitigating entity errors - *names, locations, numbers, dates, pronouns, etc.* (Kryściński, et. al, 2020, Cao, et. al, 2020, Dong, et. al, 2020, Fabbri, et. al, 2022)

Evidence

The first vaccine for Covid-19 might not be ready this year…. For reference the vaccine for Ebola took the FDA 5 years ……. be available by the end of the year.

The first vaccine for Polio took 3 years to be produced by the CBP. To produce the vaccine, scientists have to show successful human trials, then sequence the DNA of the virus.

The first vaccine for Ebola took 5 years to be produced by the FDA. To produce the vaccine, scientists have to show successful human trials, then sequence the DNA of the virus.
Goal - A general system for correcting diverse error types

- Factual Errors actually span various complex types: *entities, relations, discourse structures*

The first vaccine for Polio took 3 years to be *produced by* the CBP. To produce the vaccine, scientists have to show successful human trials, *then* sequence the DNA of the virus.

The first vaccine for Ebola took 5 years to be approved by the FDA. To produce the vaccine, scientists have to show successful human trials, *after* sequencing the DNA of the virus.
Challenges in collecting training data with diverse error types for training the Correction Model

- **Training Data: (Incorrect Text, Correct Text) Pairs**

- **Human Annotated Data**
  - **Expensive** - Long Process to read and edit text \(\text{(Pagnoni, Balachandran et. al, 2021, Min et. al, 2023)}\)
  - **Subjective** - Factuality decisions have low agreement across annotators \(\text{(Falke et al, 2019, Durmus et al, 2020)}\)

- **Synthetic Data** - Create synthetic incorrect text, are often entity oriented \(\text{(Kryściński et. al, 2020, Cao et. al, 2020, Chen et. al, 2023)}\)
Limitations with prior synthetic data

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Original sentence</th>
<th>Transformed sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paraphrasing</td>
<td>Sheriff Lee Baca has now decided to recall some 200 badges his department has handed out to local politicians just two weeks after the picture was released by the U.S. attorney’s office in support of bribery charges against three city officials.</td>
<td>Two weeks after the US Attorney’s Office issued photos to support bribery allegations against three municipal officials, Lee Baca has now decided to recall about 200 badges issued by his department to local politicians.</td>
</tr>
<tr>
<td>Sentence negation</td>
<td>Snow was predicted later in the weekend for Atlanta and areas even further south.</td>
<td>Snow wasn’t predicted later in the weekend for Atlanta and areas even further south.</td>
</tr>
<tr>
<td>Pronoun swap</td>
<td>It comes after his estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.</td>
<td>It comes after your estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.</td>
</tr>
<tr>
<td>Entity swap</td>
<td>Charlton coach Guy Luzon had said on Monday: ’Alou Diarra is training with us.’</td>
<td>Charlton coach Bordeaux had said on Monday: ’Alou Diarra is training with us.’</td>
</tr>
<tr>
<td>Number swap</td>
<td>He says he wants to pay off the $12.6million lien so he can sell the house and be done with it, according to the Orlando Sentinel.</td>
<td>He says he wants to pay off the $3.45million lien so he can sell the house and be done done with it, according to the Orlando Sentinel.</td>
</tr>
<tr>
<td>Noise injection</td>
<td>Snow was predicted later in the weekend for Atlanta and areas even further south.</td>
<td>Snow was was predicted later in the weekend for Atlanta and areas even further south.</td>
</tr>
</tbody>
</table>
## Limitations with prior synthetic data - Heuristic entity based errors

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Original sentence</th>
<th>Transformed sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prior Work</strong></td>
<td>Paraphrasing Baca has now decided to recall some 200 badges his department has handed out to local politicians just two weeks after the picture was released by the U.S. attorney’s office in support of bribery charges against three city officials.</td>
<td>Two weeks after, the U.S. Attorney’s Office issued photos to support bribery allegations against three municipal officials, Lee Baca has now decided to recall about 200 badges issued by his department to local politicians.</td>
</tr>
<tr>
<td><strong>Low coverage of diverse error types</strong></td>
<td>Sentence negation Snow was predicted later in the weekend for Atlanta and areas even further south.</td>
<td>Snow was predicted later in the weekend for Atlanta and areas even further south.</td>
</tr>
<tr>
<td><strong>Our Work</strong></td>
<td>Pronoun swap It comes after his estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.</td>
<td>It comes after his estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.</td>
</tr>
<tr>
<td><strong>Moving from entity level -&gt; Generating diverse synthetic errors at phrase/sentence level</strong></td>
<td>Entity swap Charles was told on Monday that Draghi is leaving office.</td>
<td>Charles was told on Monday: Draghi is leaving office.</td>
</tr>
<tr>
<td><strong>Moving from heuristics -&gt; Leveraging LMs to generate challenging, synthetic data</strong></td>
<td>Number swap He says he wants to pay off the $12.6 million lien so he can sell the house and be done with it, according to the Orlando Sentinel.</td>
<td>He says he wants to pay off the $3.6 million lien so he can sell the house and be done with it, according to the Orlando Sentinel.</td>
</tr>
<tr>
<td><strong>Low performance on real factual errors from stronger models</strong></td>
<td>Noise injection Snow was predicted later in the weekend for Atlanta and areas even further south.</td>
<td>Snow was was predicted later in the weekend for Atlanta and areas even further south.</td>
</tr>
</tbody>
</table>
Fava 🌱: Factuality Verification and Correction in Large LMs

**Step 1: LLM-based Generation of Synthetic Error Text**

- Factual Data → InstructLM → Synthetic Data
- Synthetic Data → InstructLM → Synthetic Incorrect Text
- Synthetic Incorrect Text → Correction Model → Correct Text

**Step 2: Training Factual Error Correction Model**

- Synthetic Incorrect Text → Correction Model → Correct Text
- (Any) Model-Generated Text → Correction Model → Revised Text

**Step 3: Correcting Model Generated Text**
Producing Factual Text as targets for training

Text: Rishi Sunak (Born 12 May 1980) is a British politician who has served as Prime Minister of the United Kingdom.

Instructions:
Paraphrase the text in News Style
Paraphrase the text in Biography Style

Diversified Output: Rishi Sunak is the current British...
Diversified Output: Rishi Sunak is an Indian-Origin....
Diversified Output: Introducing Rishi Sunak...
Inserting factual errors in factually accurate text

Instructions:
Error Definitions
Where to insert error
Edge cases to avoid

Demonstrations:
{Text, Evidence, Synthetic Output}

Text: Introducing Rishi Sunak: British politician who has served in various roles within the UK government
Evidence: Rishi Sunak (Born 12 May 1980) is a British politician who has served as Prime Minister of the United Kingdom....

Introducing Rishi Sunak: <entity>Indian</entity> politician who has served in various roles within the UK government. <unverifiable>He was an avid golfer during his graduate school days.</unverifiable>

Introducing Rishi Sunak: Indian politician who has served in various roles within the UK government. He was an avid golfer during his graduate school days.
Finetuning LM on Synthetic Training Data

**Evidence:** Rishi Sunak (born 12 May 1980) is a British politician…

**Text:** Introducing Rishi Sunak: Indian politician who has served in various roles within the UK government. He was an avid golfer during his graduate school days.

Introducing Rishi Sunak: <entity>British</entity> politician who has served in various roles within the UK government. <unverifiable><mark>He was an avid golfer during his graduate school days. </mark></unverifiable>
Evidence: Harry Potter, fictional character, a boy wizard created by British author …

Text: Harry Potter is a series of seven fantasy novels written by American author J. K. Rowling. The novels were written while J.K.Rowling frequented a coffee shop in Dublin.

Factuality Verifier+Reviser
Finetuned LLM

Harry Potter is a series of seven fantasy novels written by
<entity>
<insert>British</insert>
<delete>American</delete>
</entity> author J.K. Rowling.
<unverifiable>
<mark>The novels were written while J.K.Rowling frequented a cafe in Dublin.
</mark>
</unverifiable>
Experiment Settings

- Data Generation Model - ChatGPT
- Finetuning Model - Llama 2 7B
- Retriever - Contriever-MSMARCO (Izacard et al., 2021)

Generated Dataset Statistics
- Number of Instances - 35,074
- Avg. number of errors per passage - 3.1
Evaluation Setup

- **Task-1: Error Detection**
  - Accuracy on Human-Annotated Error Type Data
  - Data: Open Assistant, Instruction Following Queries, WebNLG

- **Task-2: Error Correction**
  - Wikipedia Entity Biography Generation ([Min et al. 2023](#))
  - FactScore ([Min et al. 2023](#)) - measure precision w.r.t. to facts from Wikipedia
## Error Type Detection Results

### ChatGPT

<table>
<thead>
<tr>
<th>Method</th>
<th>Type Level Acc</th>
<th>Binary Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChatGPT+FewShot Refine</td>
<td>18.8</td>
<td>50.1</td>
</tr>
<tr>
<td>Retrieval + ChatGPT+FewShot Refine</td>
<td>24.4</td>
<td>64.8</td>
</tr>
<tr>
<td>Fava</td>
<td>46.5</td>
<td>78.2</td>
</tr>
</tbody>
</table>

### LLama

<table>
<thead>
<tr>
<th>Method</th>
<th>Type Level Acc</th>
<th>Binary Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChatGPT+FewShot Refine</td>
<td>24.1</td>
<td>68.4</td>
</tr>
<tr>
<td>Retrieval + ChatGPT+FewShot Refine</td>
<td>27.8</td>
<td>72.8</td>
</tr>
<tr>
<td>Fava</td>
<td>46.5</td>
<td>80.6</td>
</tr>
</tbody>
</table>
Error Type Detection Results

Fine-Grained Type Level Performance

- **CGPT**
- **R+CGPT**
- **FAVA**

![Graph showing F1 scores for different error types]

- **Error Type**: Entity Contradiction, Relation Contradiction, Contradictor Sentence, Invented Concept, Subjective Statement, Unverifiable Statement

**Factual Error Correction**
Error Correction Results

<table>
<thead>
<tr>
<th>Method</th>
<th>ChatGPT</th>
<th>Alpaca-7B</th>
<th>Alpaca-13B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Model Generation (NoEdit)</td>
<td>66.7</td>
<td>38.8</td>
<td>42.5</td>
</tr>
<tr>
<td>ChatGPT+FewShot Refine</td>
<td>58.6</td>
<td>37.9</td>
<td>42.0</td>
</tr>
<tr>
<td>Retrieval + ChatGPT+FewShot Refine</td>
<td>62.7</td>
<td>39.2</td>
<td>43.9</td>
</tr>
<tr>
<td>LLama+FewShot Refine</td>
<td>52.6</td>
<td>18.6</td>
<td>22.7</td>
</tr>
<tr>
<td>Retrieval + LLama+FewShot Refine</td>
<td>58.7</td>
<td>32.2</td>
<td>48.6</td>
</tr>
<tr>
<td>Fava</td>
<td><strong>70.0 (+3.3)</strong></td>
<td><strong>51.8 (+9.3)</strong></td>
<td><strong>43.2 (+3.3)</strong></td>
</tr>
</tbody>
</table>
Summary

● Fava - Error Verification and Correction for Open-Ended Generation
  ○ Retrieval-Augmented Model for verifying+correcting model generated text
  ○ Model trained to “mark” incorrect text for deletion and “insert” suggestions for replacement

● Leveraging Instruction Tuned models for synthetic data generation
  ○ Using LLMs to produce fine-grained, diverse adversarial data for training
  ○ Flexible, Controllable and Customizable process enabling better training data distribution

● Applicable across diverse error categories
  ○ Generated training data includes diverse examples of errors
  ○ Diverse, high-quality data generation helps error correction across multiple models and error categories
Open Questions and Future Work

- Improving Signals and Objectives for Training
- Incorporating Diverse Sources of Reliable Knowledge
- Safety and Reliability for Multimodal, Continual Systems
Future Work - Training Signals and Methods for Reliability

- Current pre-training methods encourage plausible language generation and collecting preference data for diverse aspects of reliability is under-explored.
- Need better signals of attributable and factual text for training, fine-grained rewards for encouraging nuanced aspects of factuality.

Improving alignment with factual data using RLHF and fine-grained preference

Augmenting self-supervised learning/pre-training to teach plausible+accurate language generation
Diverse Sources of Reliable Knowledge

- Sources of world knowledge and facts are diverse with varying levels of veracity - news, books, encyclopedias, tabloids, magazines, textbooks and more!
- Need to aggregate knowledge from multiple sources by taking into account their reliability for complex fact-checking

Ongoing Work: Augmenting Models with External Tools for Fact-Checking

CooK: Language Models with Modular and Collaborative Knowledge (Feng, Shi, Bai, Balachandran, et. al, 2023)
Reliability for Multimodal Settings

Describe the image.

I can see the image now. It shows a group of smartphones sitting on top of each other on a table. There are four smartphones in total, and they are all different colors. The top smartphone is black, the second one is white, the third one is green, and the bottom one is blue. The smartphones are all facing the same direction, and they are all turned on. The screens of the smartphones are all blank.

The smartphones are sitting on a wooden table. The table is brown and it has a few scratches on it. There is a white piece of paper on the table next to the smartphones. The paper is blank.
Reliability for Multimodal Settings

- Multi-Modal pretraining incorporates various sources of knowledge - images, videos, text, speech
- Fundamental research on knowledge, factuality, evaluation and mitigation in context of multimodal models is necessary
Summary and Takeaways

- Studying, Detecting and Mitigating Factual Errors is a **challenging problem that needs urgent attention** from research, modeling and application perspective.

- Factual Errors and Hallucinations can manifest in variety of different ways highlighting the need for more generalizable solutions to address factuality.

- Some initial work on studying and mitigating factual errors - FactKB, FAVA.

- The **challenges with factuality is getting larger and more complex** with development of multimodal AI systems and growing applications of AI systems.
Thank you and Questions

- Studying, Detecting and Mitigating Factual Errors is a challenging problem that needs urgent attention from research, modeling and application perspective

- Factual Errors and Hallucinations can manifest in variety of different ways highlighting the need for more generalizable solutions to address factuality

- Some initial work on studying and mitigating factual errors - FactKB, FAVA

- The challenges with factuality is getting larger and more complex with development of multimodal AI systems and growing applications of AI systems

Summary and Takeaways

https://github.com/BunsenFeng/FactKB

https://huggingface.co/bunsenfeng/FactKB

Email: vbalacha@cs.cmu.edu