Building Reliable LLMs
Evaluating and Mitigating Factual Inconsistencies in Language Generation

Vidhisha Balachandran
vbalacha@cs.cmu.edu

December 1, 2023

Carnegie Mellon University
Language Technologies Institute

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!

The Economist

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

POLITICO

THE FIFTY

More schools want your kids to use ChatGPT. Really.

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

REUTERS®

How will leveraging AI change the future of legal services?

Healthcare IT News

NYU Langone Health LLM can predict hospital readmissions

Microsoft Research Blog

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

The Verge

Bing, Bard, and ChatGPT: How AI is rewriting the internet

But Pretrained Large LMs still generate a *variety* of Factual Errors

V Write a detailed biography of Quico Canseco

Quico Canseco, whose full name is Francisco Javier Canseco Jr., is a Mexican-American lawyer and former politician who served as a member of the United States House of Representatives. Born on July 30, 1949, in Laredo, Texas, Canseco's life and career have been marked by his involvement in law, business, and public service.

After completing his undergraduate studies at the University of Texas at Austin, where he earned a Bachelor of Business Administration degree, Canseco went on to pursue a Juris Doctor degree from St. Mary's University School of Law in San Antonio, Texas. He then began his legal career, specializing in international trade law and business litigation.

Generating wrong entities and attributes

```
Patient's facts:
- 20 year old female
- with a history of anerxia nervosa and depression
- blood pressure 100/50, pulse 50, height 5'5''
- referred by her nutrionist 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 nu
tritionist 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 (100/50) and bradycardia
(pulse 50). She also reports feeling depressed and hopeless. She is at risk of serious medical compli
cations, 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

Input Sentence	Model Generated Paraphrase
Two tables next to each other along with laptops	two people sitting on the beach with their laptops
a city street line with very tall buildings	a city street with several signs on the street

Generating ungrounded entities

PA summarise this article https://www.nytimes.com/2023/03/11/technology/chatgpt-prompts-to-avoid-content-filters.html

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.

Hallucinating entire content

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

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

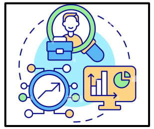
Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey (*Balachandran*, et, al. 23)

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!

GIZMODO

CNET Is Reviewing the Accuracy of All Its AI-Written Articles After Multiple Major Corrections

Big surprise: CNET's writing robot doesn't know what it's talking about.

UNIVERSITY OF ALBERTA

LIBRARY

I'm having trouble accessing an article suggested by ChatGPT. Can you help?

nature

ARTIFICIAL INTELLIGENCE

Research Summaries Written by AI Fool Scientists

Scientists cannot always differentiate between research abstracts generated by the AI ChatGPT and those written by humans

AP

Lawyers submitted bogus case law created by ChatGPT. A judge fined them $5,000

The Washington Post

Democracy Dies in Darkness

A news site used AI to write articles. It was a journalistic disaster.

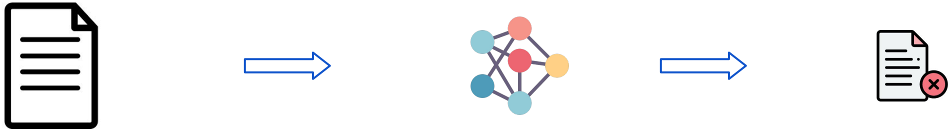
The tech site CNET sent a chill through the media world when it tapped artificial intelligence to produce surprisingly lucid news stories. But now its human staff is writing a lot of corrections.

TECH

Think twice before using ChatGPT for help with homework

This new AI tool talks a lot like a person — but still makes mistakes

Factual Accuracy of Model Generated Text

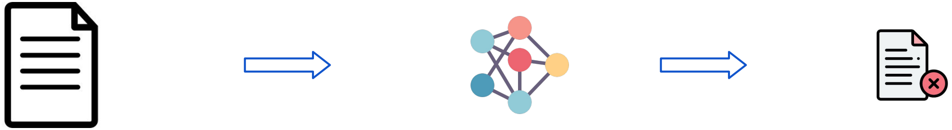


The New York Times is ending its Covid data-gathering operation. The Times will continue to publish its Covid tracking ……. impact of the virus on communities.

Source Document

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

Summary w/ Factual Errors



Explain the events in Abraham Lincoln's life in detail.

Prompt/Instruction

Abraham Lincoln was born on **March 3, 1800**, in a log cabin in Hardin County (now LaRue County), **Indiana**….

Answer w/ Factual Errors

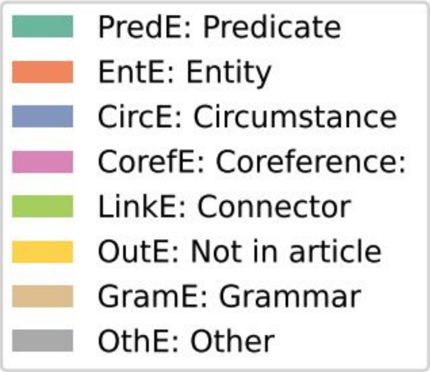
Factual Errors in Summarization vary across Datasets and Models

- 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

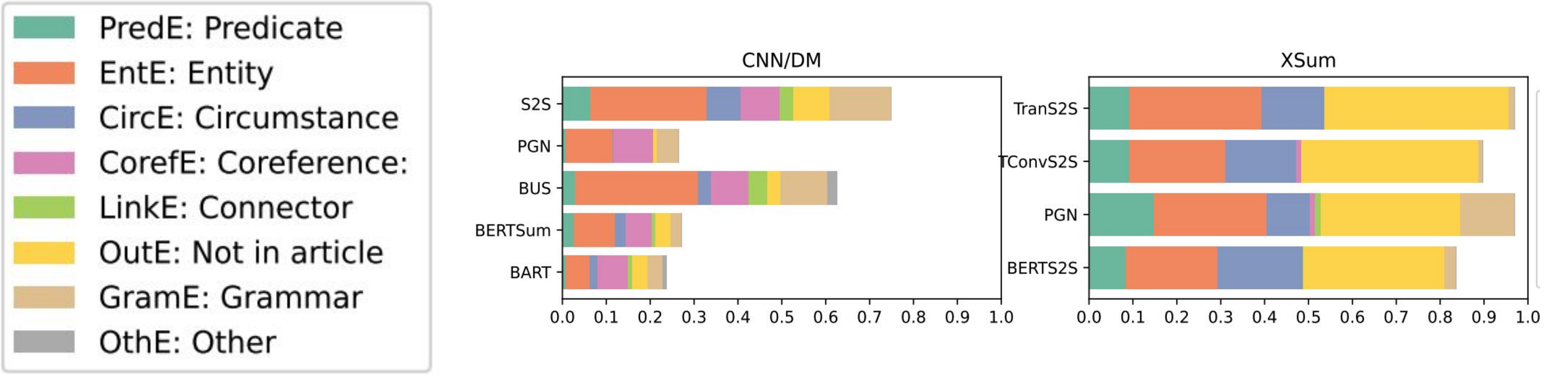
Content Verifiability Errors
Errors out of article scope



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

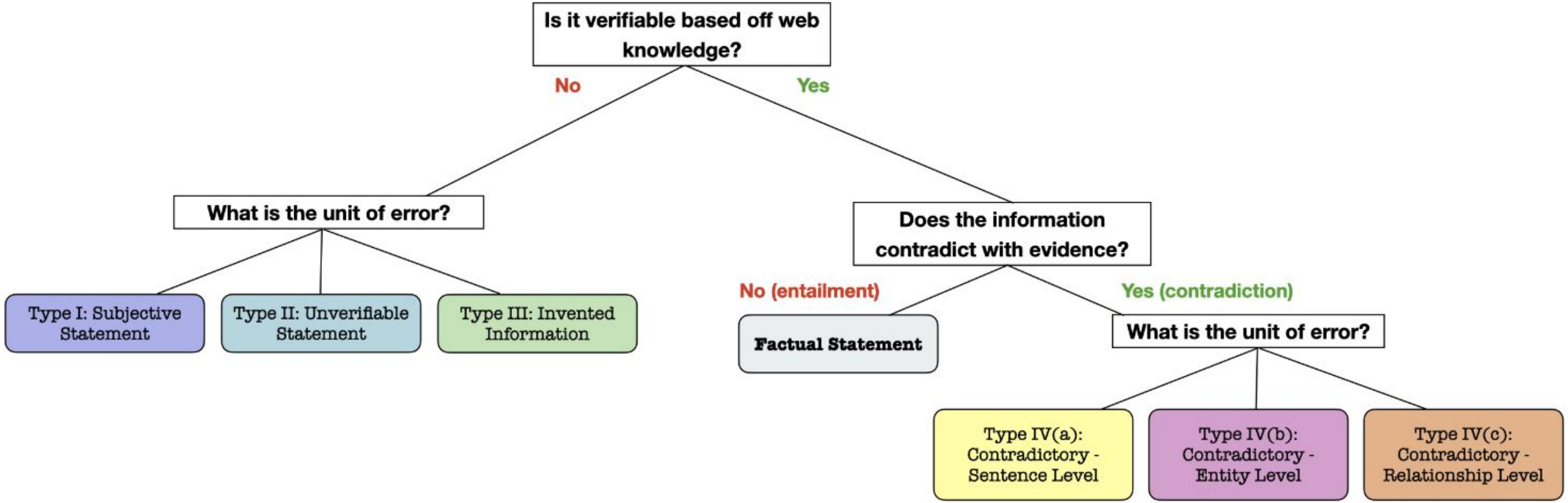
- Summaries generated by the same models consist of different error distributions over different datasets (Pagnoni, *Balachandran*, et. al, 2021, Goyal, et al. 2023)
- Error distribution can vary among models within the same category



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

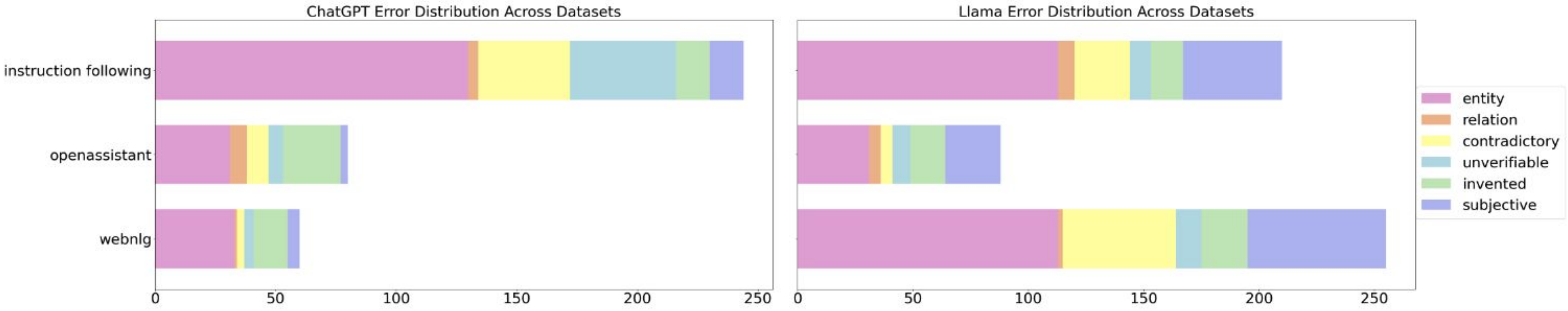
- Powerful LLMs like GPT models, LLama models produce more complex factual issues - invented concepts, unverifiable content, wrong temporal relations

Type	Example	ChatGPT	Llama2
Subjective	Lionel Messi is **the best soccer player in the world.**	12.82%	8.86%
Invented	**Messi is also famous for his discovery of the famous airplane kick technique.**	5.13%	22.97%
Unverifiable	**In his free time, Messi enjoys singing songs for his family.**	14.74%	5.06%
Contradictory	**Messi has yet to gain captaincy for the Argentina national football team.**	14.74%	14.10%
Entity	Lionel Andrés Messi was born on June ~~12~~ **24**, 1987.	49.36%	46.47%
Relation	Lionel Messi ~~acquired~~ **was acquired by** Paris Saint-Germain.	3.21%	2.53%

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



FAVA: Understanding and Correcting Hallucinations in Large Language Models (forthcoming Mishra, *Balachandran* et. al, 2023)

Generalizable Factuality Evaluation

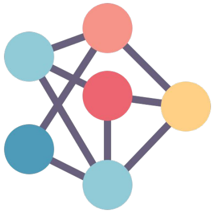
FactKB: Generalizable Factuality Evaluation using Language Models Enhanced with Factual Knowledge (Feng, *Balachandran*, et. al, *EMNLP 2023)*



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 ….



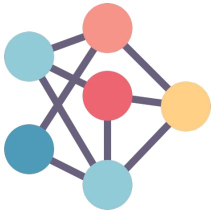
Error Detector

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

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 ….



Error Detector

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

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

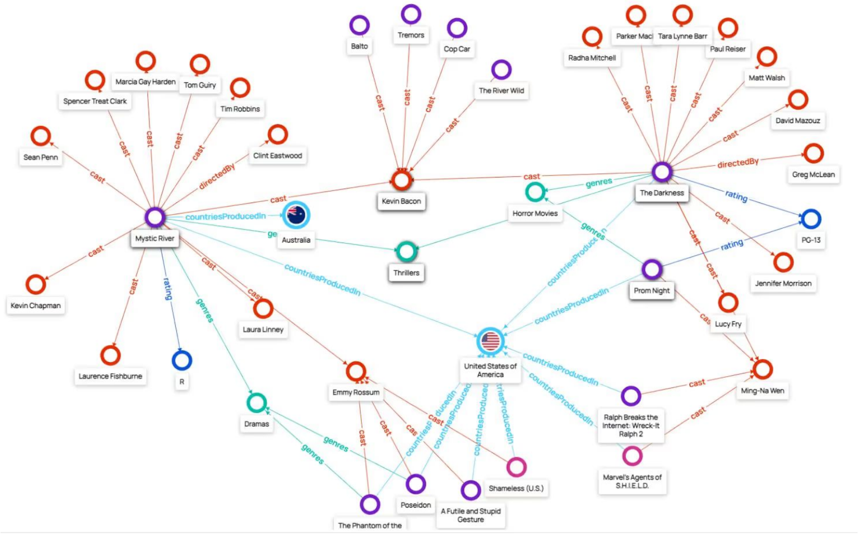
- Human Annotated Data
  - Expensive - Long Process to read and label summaries (Pagnoni, *Balachandran* et. al, 2021, Min et. al, 2023)
  - Subjective - Factuality decisions have low agreement across annotators (Falke et al, 2019, Durmus et al, 2020)

- 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

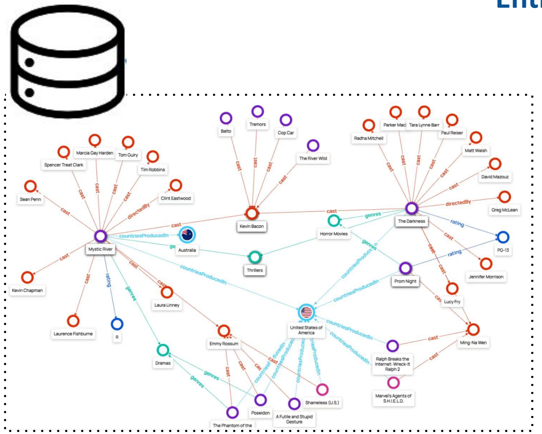
Structured KB Facts for Diverse Entity Knowledge

- 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)



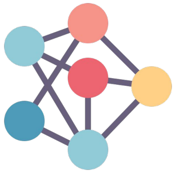
FactKB: Leveraging KB Facts to Pretrain LMs for Factuality Detection

Knowledge Base



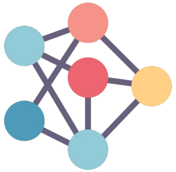
Entity-Oriented Pretraining Objectives

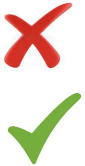
1. **Entity Wiki**
2. **Evidence Extraction**
3. **Knowledge Walk**



Step1: Pretrain LM on Structured KB Facts

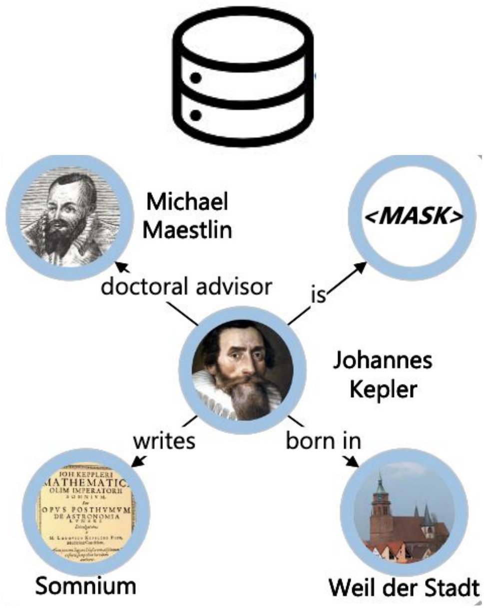
Document: …
Generated Summary: …





Step2: Finetune LM on Human-Annotated Data

Construct Statements from KB Facts



Use KB Facts to construct Surface form Statements

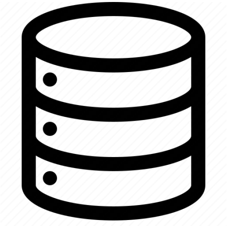
Johannes Keppler doctoral advisor Michael Maestin

Johannes Keppler born in Well der Stadt on 27 December 1571

Johannes Keppler was an astronomer, mathematician, physicist

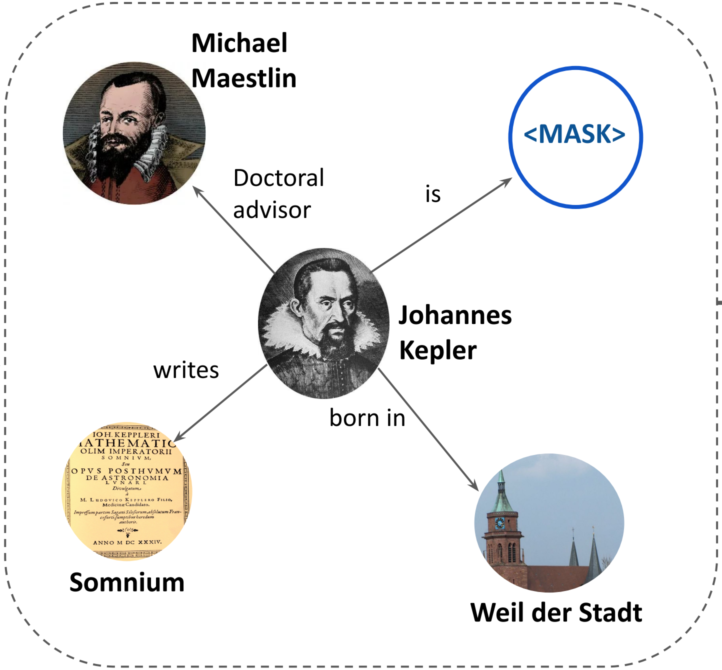
Somnium written by Johannes Keppler

Pretraining Objective 1 - Entity Wiki



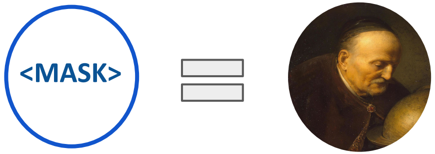
Knowledge Base

Extract Structured Facts



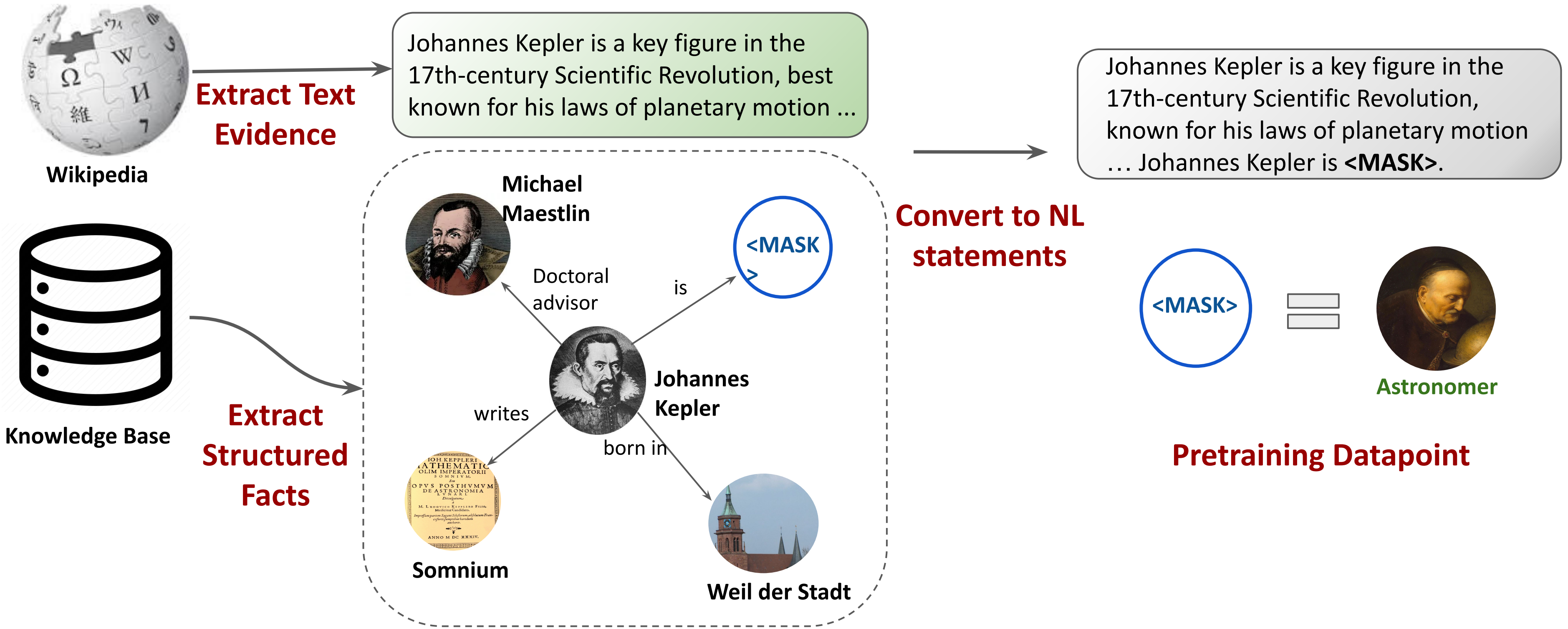
Convert to NL statements

Kepler doctoral advisor Michael Maestin. Kepler is <MASK>. Kepler born in Weil der Stadt. Kepler writes Somnium

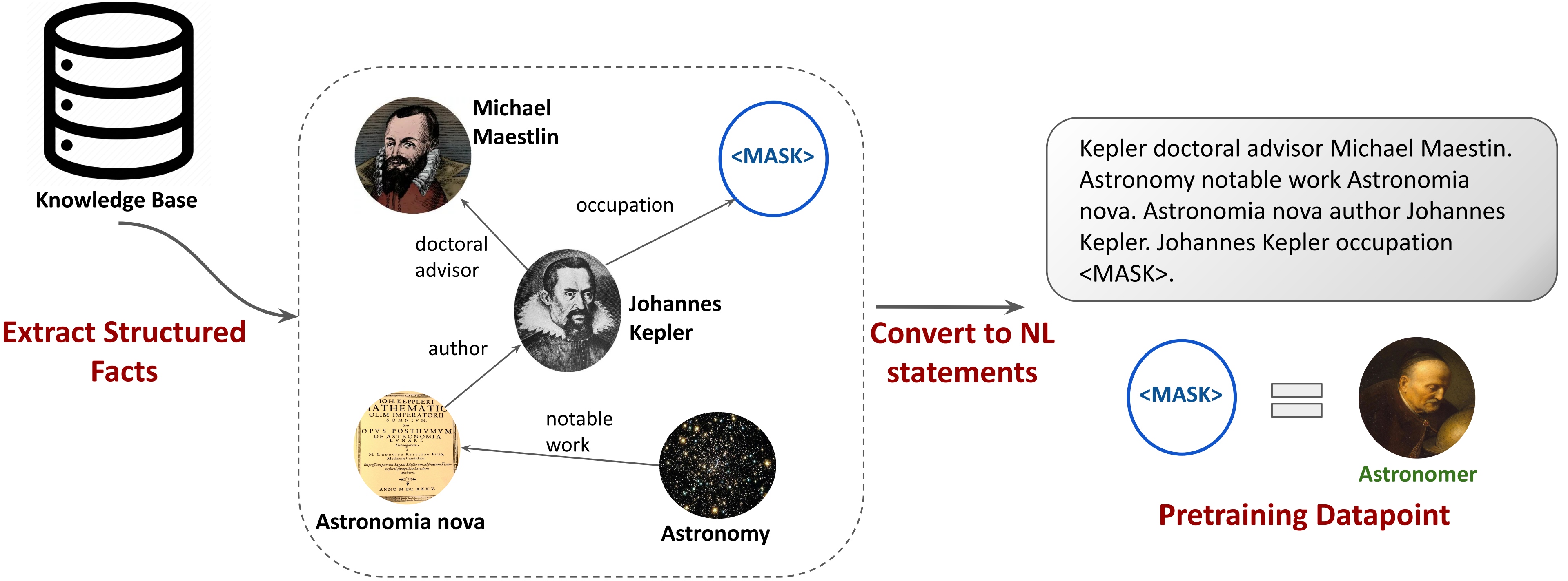


Pretraining Datapoint

Pretraining Objective 2 - Evidence Extraction



Pretraining Objective 3 - Knowledge Walk



Pretraining Corpora Details

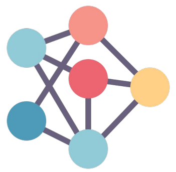
Factuality Pretraining	Corpus Size Bound	# Tokens
Entity Wiki	$\propto \lvert\mathcal{E}\rvert$	5.4M
Evidence Extraction	$\propto \lVert A\rVert_0$	12.2M
Knowledge Walk	$\propto \lvert\mathcal{E}\rvert(\frac{\lVert\mathcal{A}\rVert_0}{\lvert\mathcal{E}\rvert})^k$	2.7M

Finetuning FactKB for Factual Error Detection

Training Document	Model Generated Summary	Label
The first vaccine for Covid-19 …….. ready this year, although clinical trials have already started. For reference the vaccine for Ebola took ……..	Vaccine for Ebola is unlikely to be ready this year.	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)

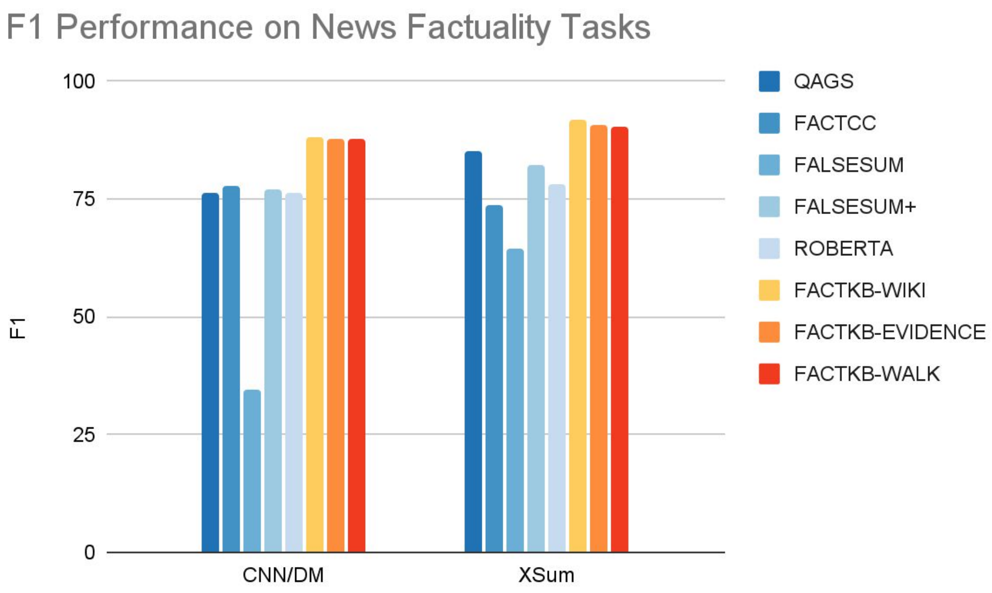
Evaluation Setup

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

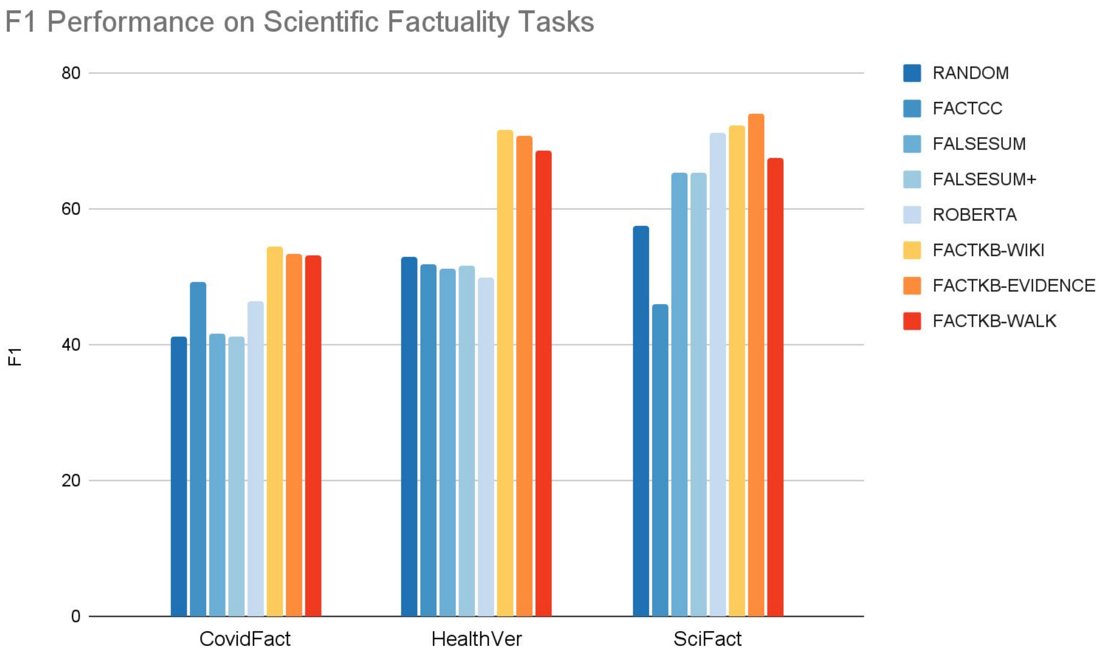
- Zero-Shot Scientific Fact-Checking Evaluation:
  - CovidFact (Saakyan et al., 2021)
  - HealthVer (Sarrouti et al., 2021)
  - SciFact (Wadden et al., 2020)

- Baselines:
  - QA Based (Wang et al., 2020)
  - Entailment Based (Krysci´nski et al., 2020, Utama et al., 2022)
  - Roberta on FactCollect Baseline

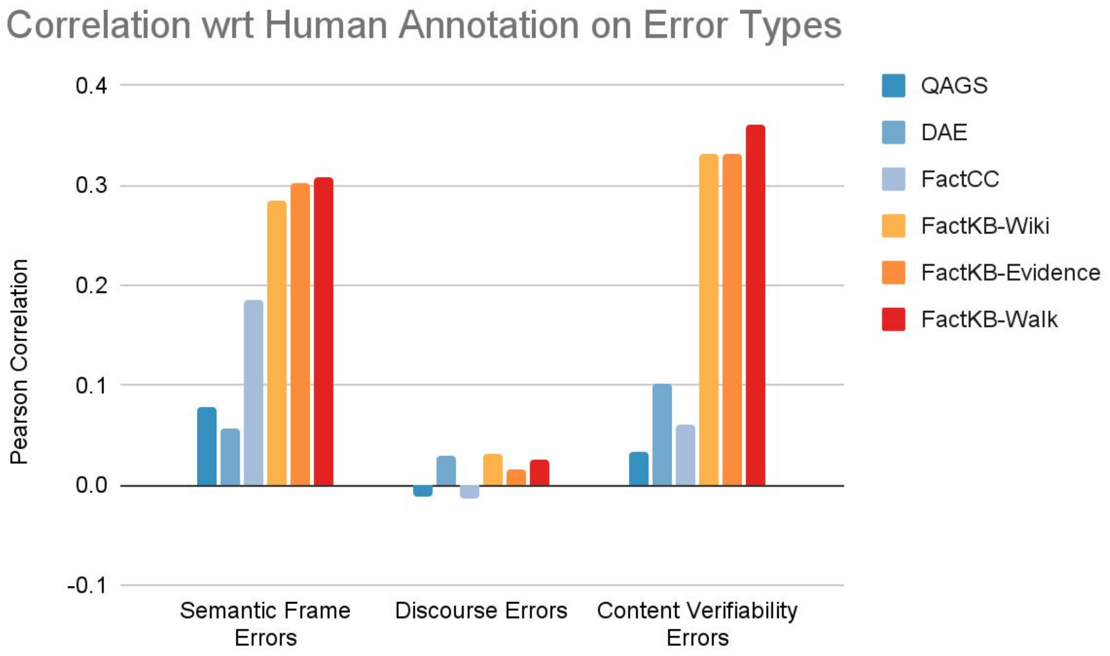
FactKB performance on News Domain



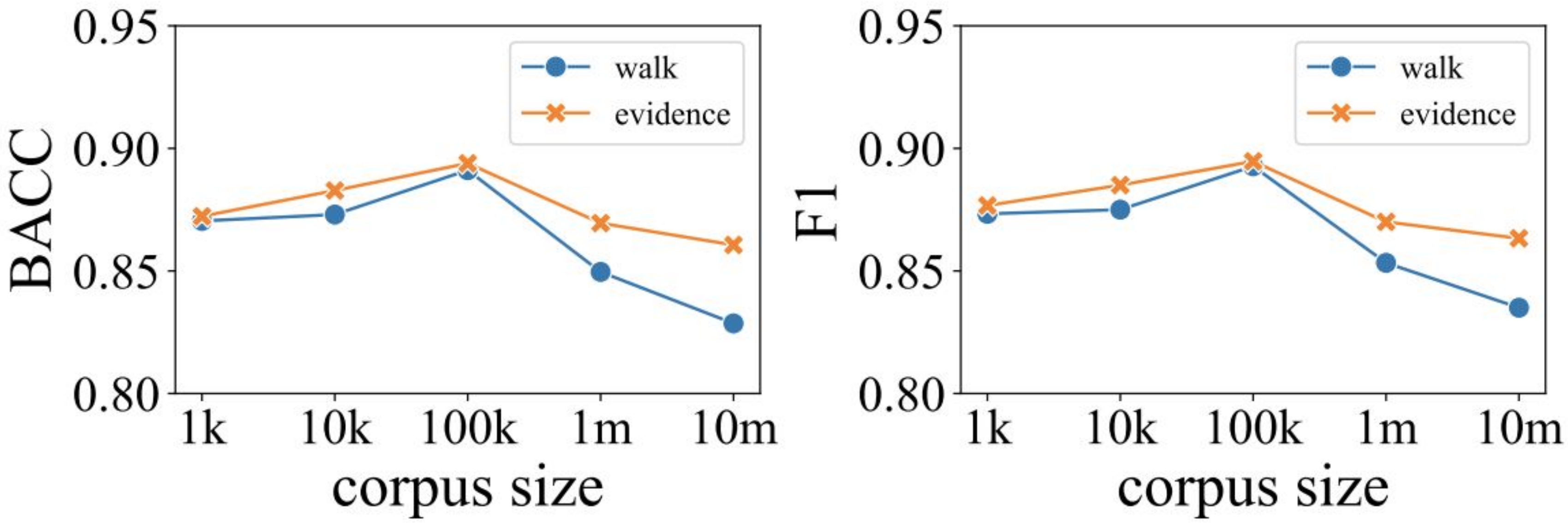
FactKB performance on Scientific Literature Domain



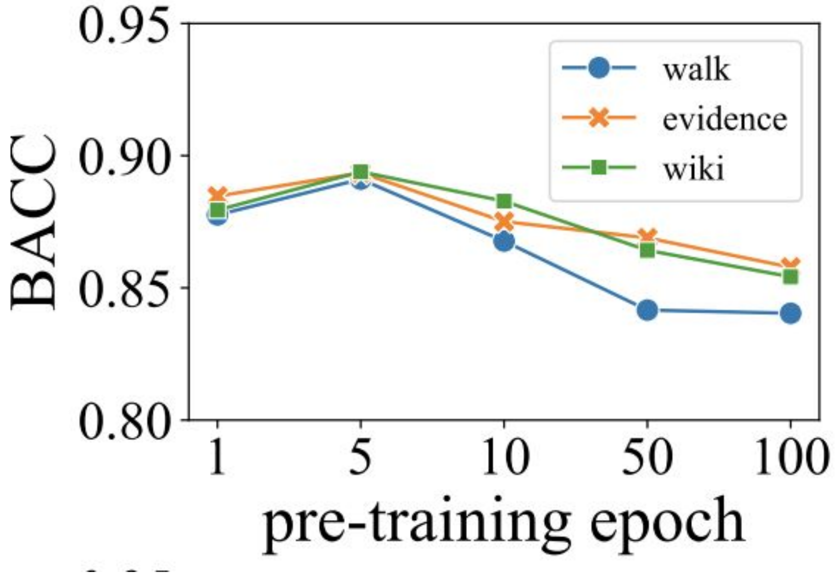
FactKB performance across error types

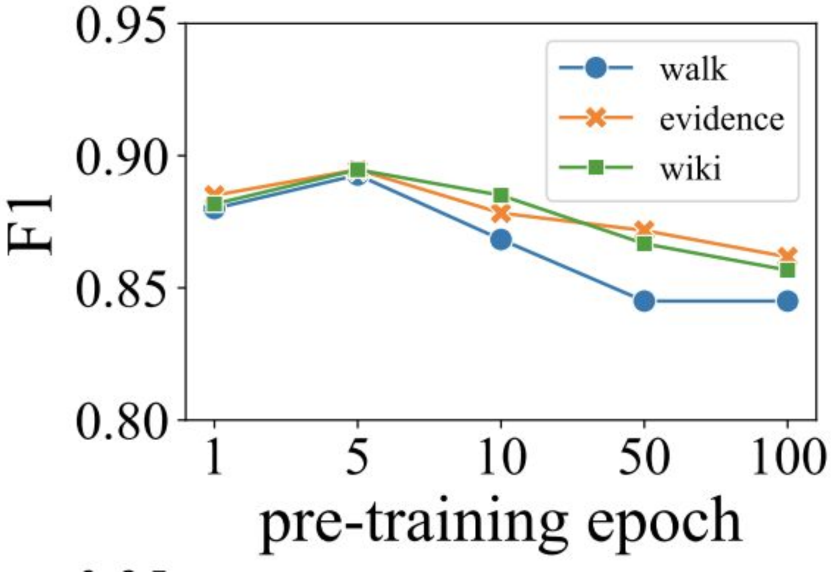


Pretraining Corpus Size effect on Performance



Pretraining Corpus Size effect on Performance





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

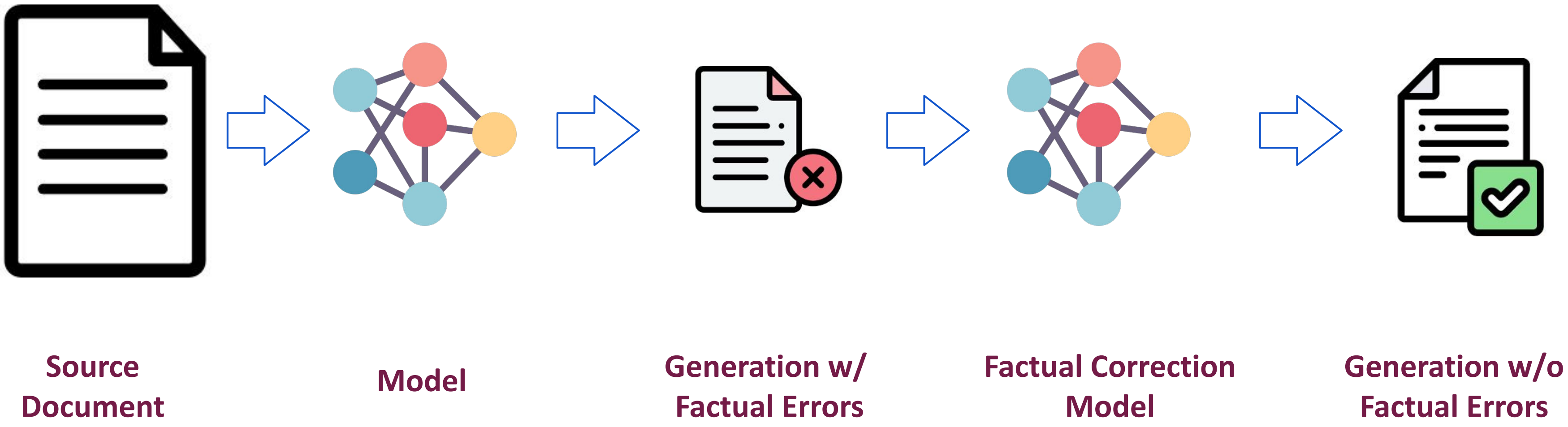
Understanding Factual Error Types and Correcting Diverse Errors

FAVA: Understanding and Correcting Hallucinations in Large Language Models *(*Mishra*, Balachandran*, et. al*, Forthcoming)*






Post-Editing to Correct Factual Errors



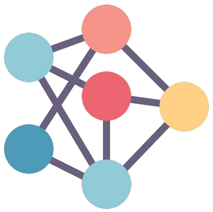
Goal - A general system for correcting diverse error types

- 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.



Correction Model

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.

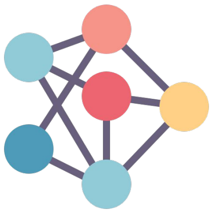
Goal - A general system for correcting diverse error types

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

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.



Correction Model

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 (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 text, are often entity oriented (Kryściński et. al, 2020, Cao et. al, 2020, Chen et. al, 2023)

Limitations with prior synthetic data

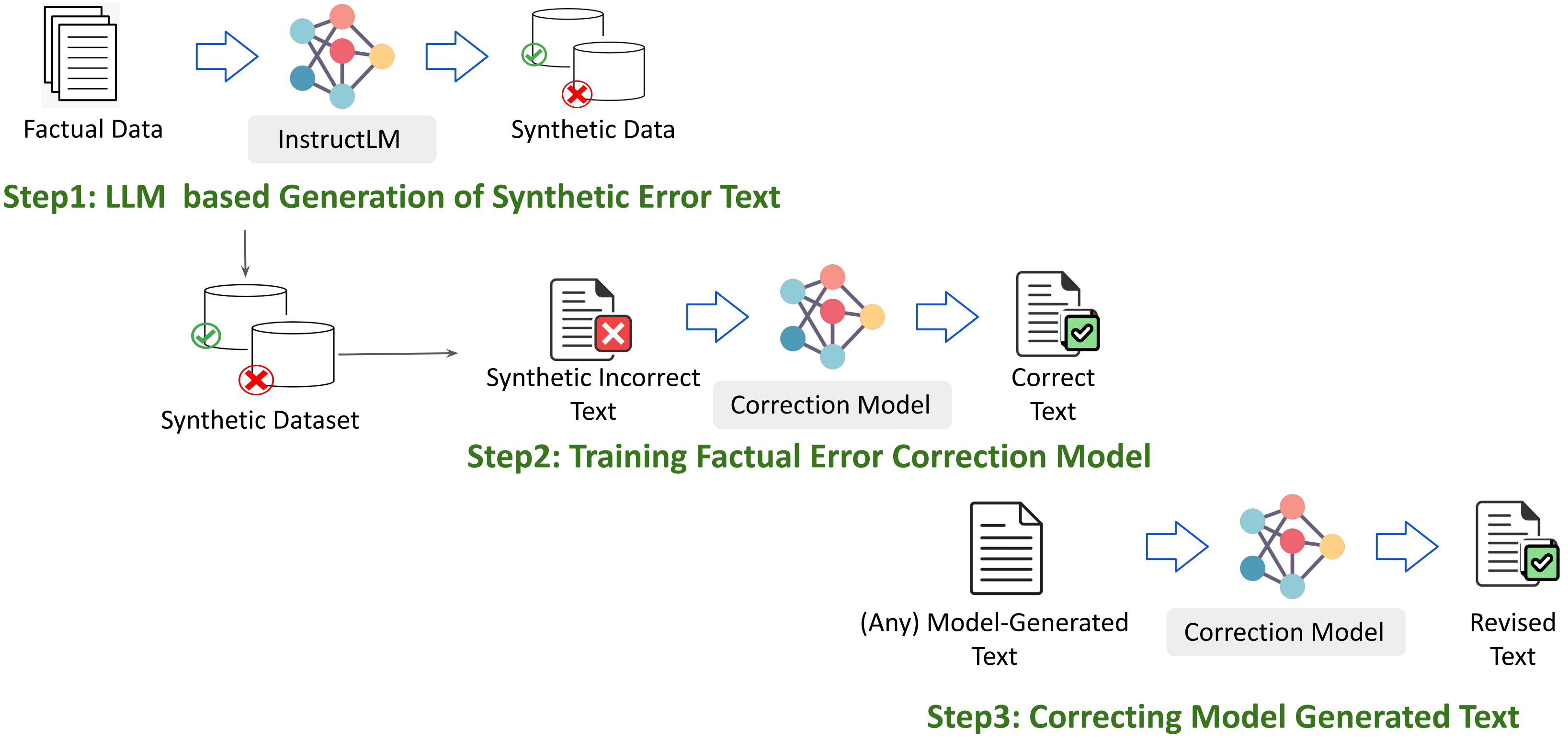
Transformation	Original sentence	Transformed sentence
Paraphrasing	Sheriff Lee Baca has now decided to recall some 200 badges his department has handed out to local politicians just two weeks after the picture was released by the U.S. attorney's office in support of bribery charges against three city officials.	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.
Sentence negation	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow wasn't predicted later in the weekend for Atlanta and areas even further south.
Pronoun swap	It comes after his estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.	It comes after your estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.
Entity swap	Charlton coach Guy Luzon had said on Monday: 'Alou Diarra is training with us.'	Charlton coach Bordeaux had said on Monday: 'Alou Diarra is training with us.'
Number swap	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.	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.
Noise injection	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow was was predicted later in the weekend for Atlanta and areas ~~even~~ further south.

Evaluating the Factual Consistency of Abstractive Text Summarization (Kryściński et al, 20)
Factual Error Correction for Abstractive Summarization Models (Cao et al, 21)

Limitations with prior synthetic data - Heuristic entity based errors

Prior Work	Our Work
Low coverage of diverse error types	Moving from entity level -> Generating diverse synthetic errors at phrase/sentence level
Low performance on real factual errors from stronger models	Moving from heuristics -> Leveraging LMs to generate challenging, synthetic data

Fava 🌱: Factuality Verification and Correction in Large LMs



Producing Factual Text as targets for training

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

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

Data-Generation - Instruction Tuned Model

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….

Data-Generation - Instruction Tuned Model

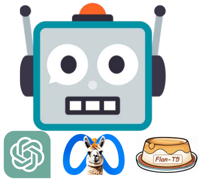
Introducing Rishi Sunak: **<entity>** **<delete>British</delete>** **<insert>Indian</insert>** **</entity>** politician who has served in various roles within the UK government. **<unverifiable>** **</insert>He was an avid golfer during his graduate school days.</insert>** **</unverifiable>**

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…

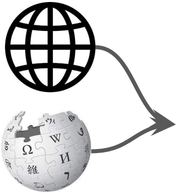
Text: Introducing Rishi Sunak: **Indian** politician who has served in various roles within the UK government. **He was an avid golfer during his graduate school days.**



Instruction-Tuned LLM

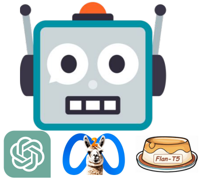
Introducing Rishi Sunak: **<entity>** **<insert>British</insert>** **<delete>Indian</delete>** **</entity>** politician who has served in various roles within the UK government. **<unverifiable>** **<mark> He was an avid golfer during his graduate school days. </mark>** **</unverifiable>**

Inference - applying Fava 🌱 on model generated text



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

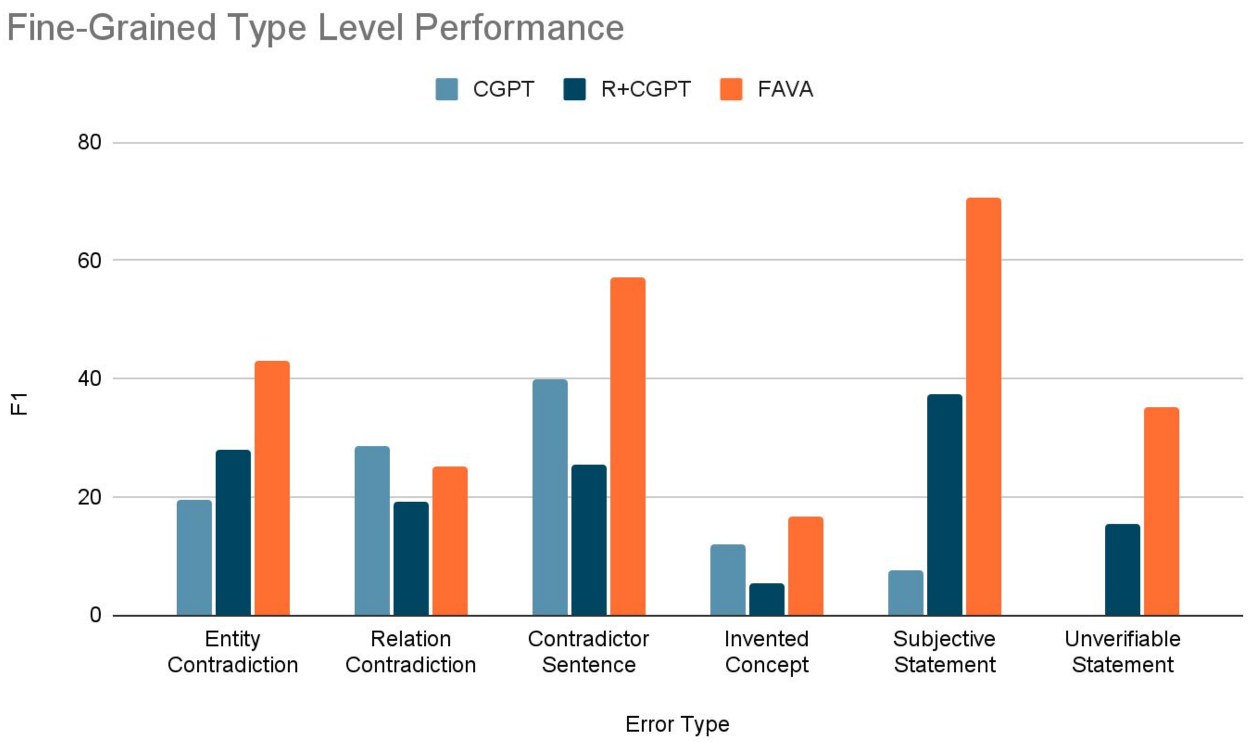
ChatGPT

Method	Type Level Acc	Binary Acc
ChatGPT+FewShot Refine	18.8	50.1
Retrieval + ChatGPT+FewShot Refine	24.4	64.8
Fava	**46.5**	**78.2**

LLama

Method	Type Level Acc	Binary Acc
ChatGPT+FewShot Refine	24.1	68.4
Retrieval + ChatGPT+FewShot Refine	27.8	72.8
Fava	**46.5**	**80.6**

Error Type Detection Results



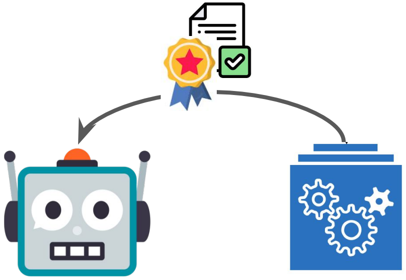
Error Correction Results

Method	ChatGPT	Alpaca-7B	Alpaca-13B
Base Model Generation (NoEdit)	66.7	38.8	42.5
ChatGPT+FewShot Refine	58.6	37.9	42.0
Retrieval + ChatGPT+FewShot Refine	62.7	39.2	43.9
LLama+FewShot Refine	52.6	18.6	22.7
Retrieval + LLama+FewShot Refine	58.7	32.2	48.6
Fava	**70.0** **(+3.3)**	**51.8** **(+9.3)**	**43.2** **(+3.3)**

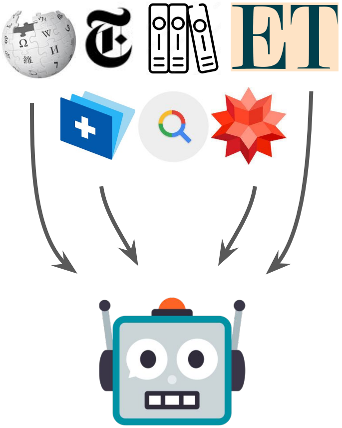
Summary

- 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

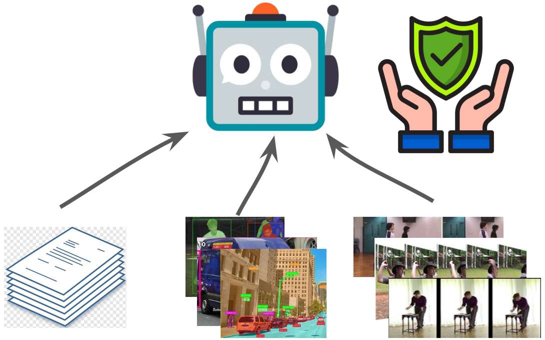
Open Questions and Future Work



Improving Signals and Objectives for Training



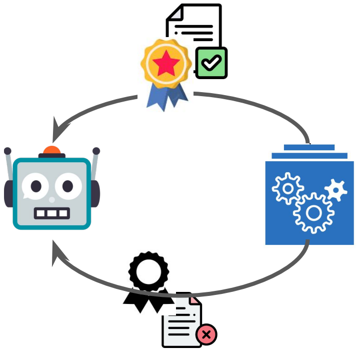
Incorporating Diverse Sources of Reliable Knowledge



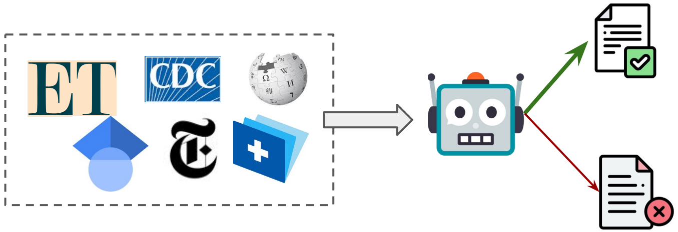
Safety and Reliability for Multimodal, Continual Systems

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.



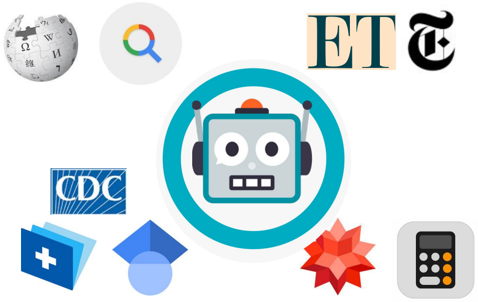
Improving alignment with factual data using RLHF and fine-grained preference



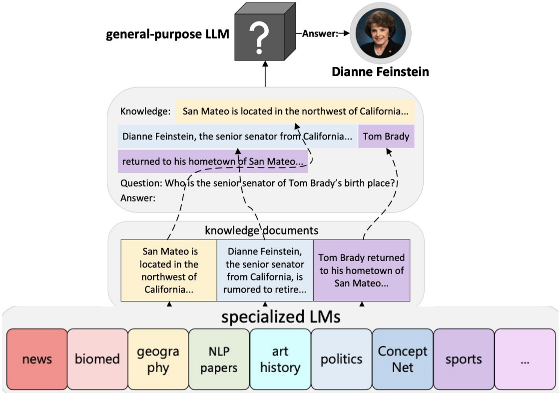
Augmenting self-supervised learning/pre-training to teach plausible+accurate language generation

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.

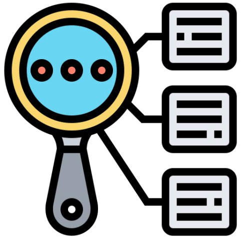
View other drafts

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



Understanding factual errors in multi-modal settings

Efficient retrieval and encoding of diverse evidence for detection and evaluation

Adapting mitigation techniques for reducing multi-modal factual errors

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

https://github.com/BunsenFeng/FactKB

https://huggingface.co/bunsenfeng/FactKB

Email : vbalacha@cs.cmu.edu


Fin