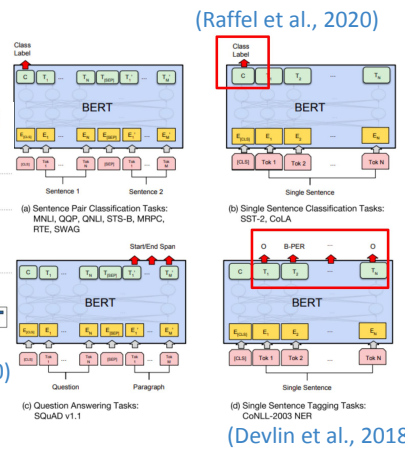
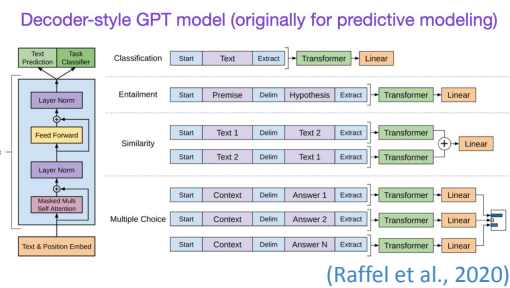
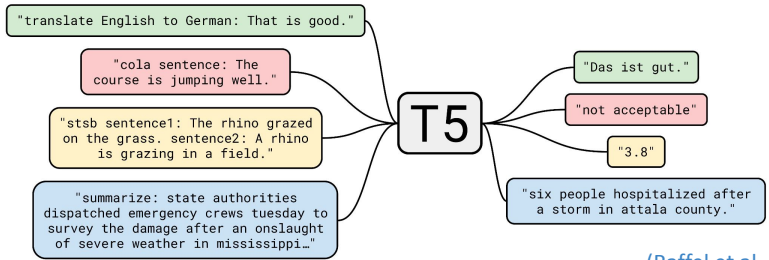
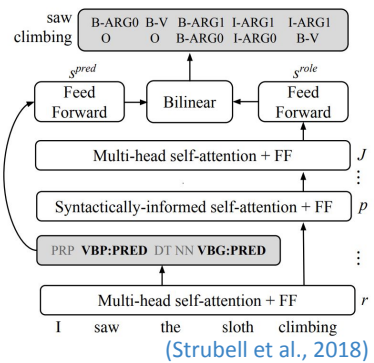
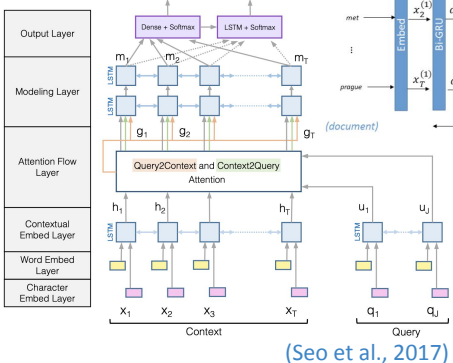
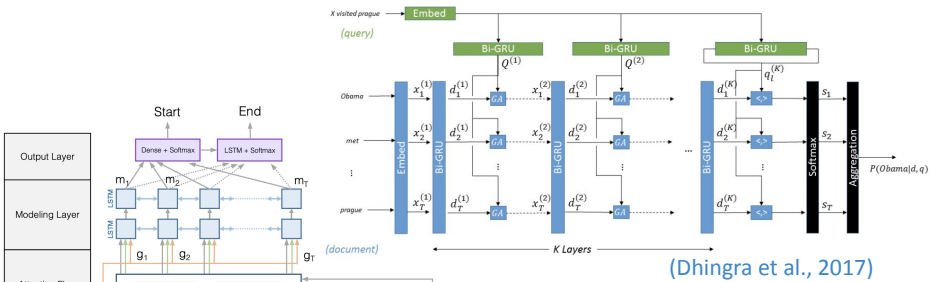


Generalizable Factual Error Correction in Model Generated Summaries

Vidhisha Balachandran
18 May, 2023

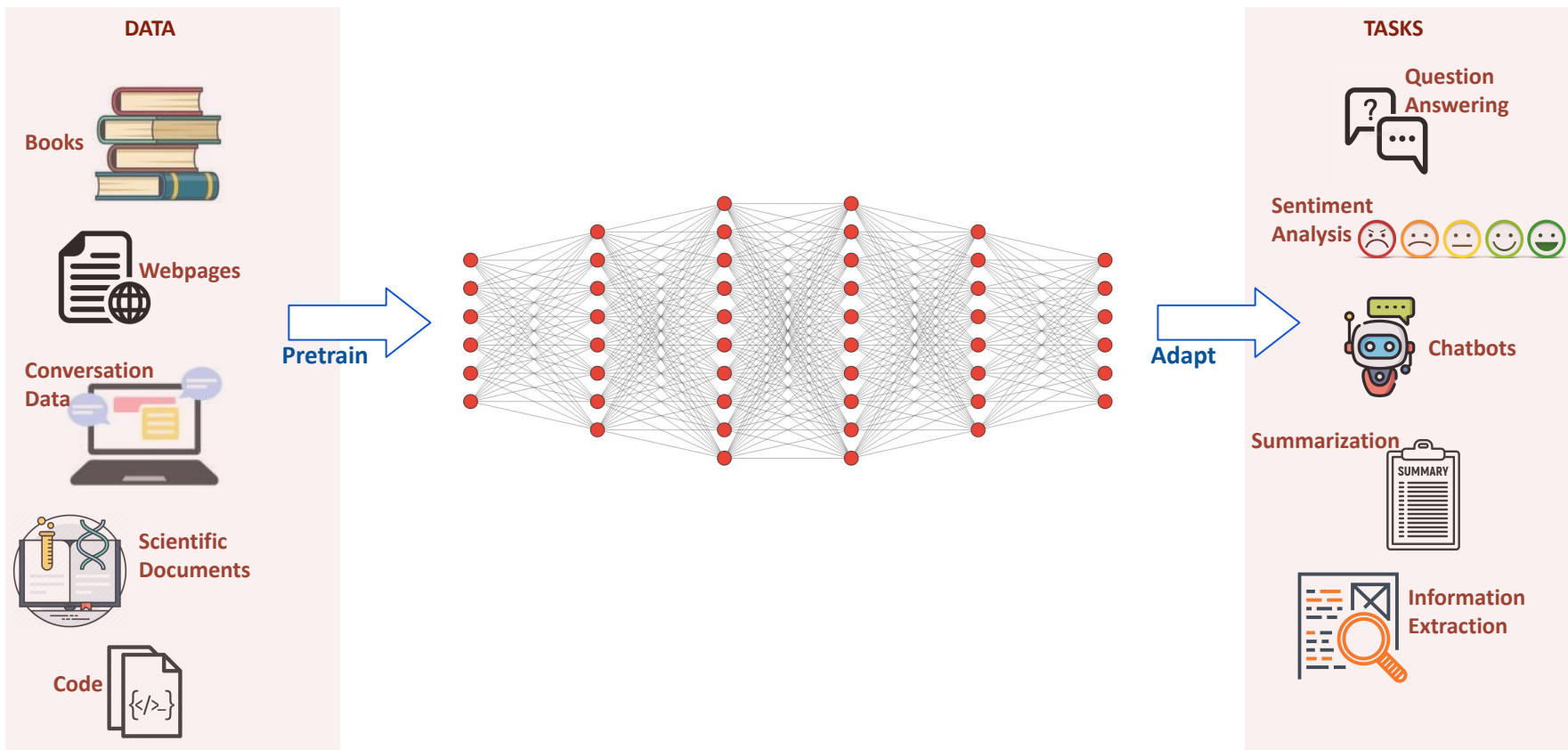
NLP Today looks very different what it was 5 years ago!



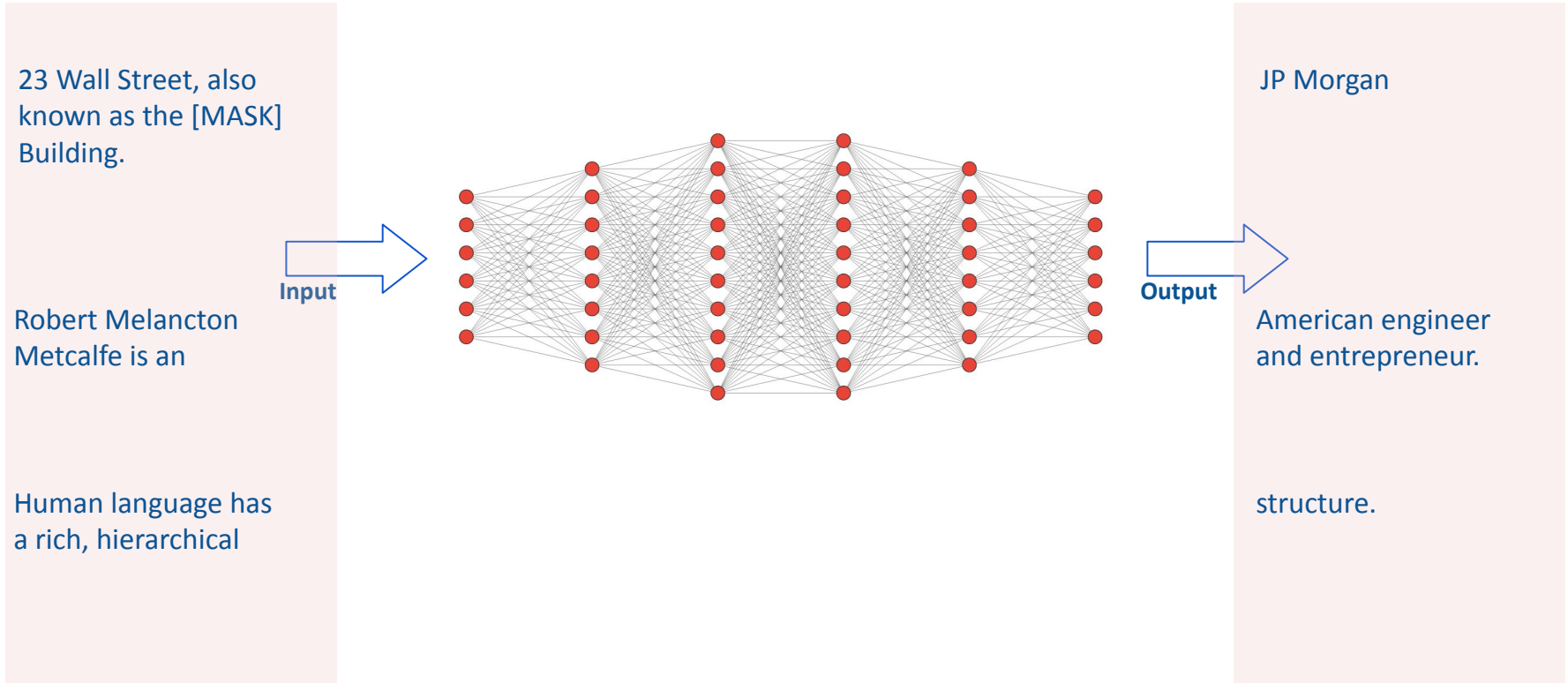
Task-Specific Architectures

General, Multi-Task Architectures

They are pretrained on large, diverse sources of data

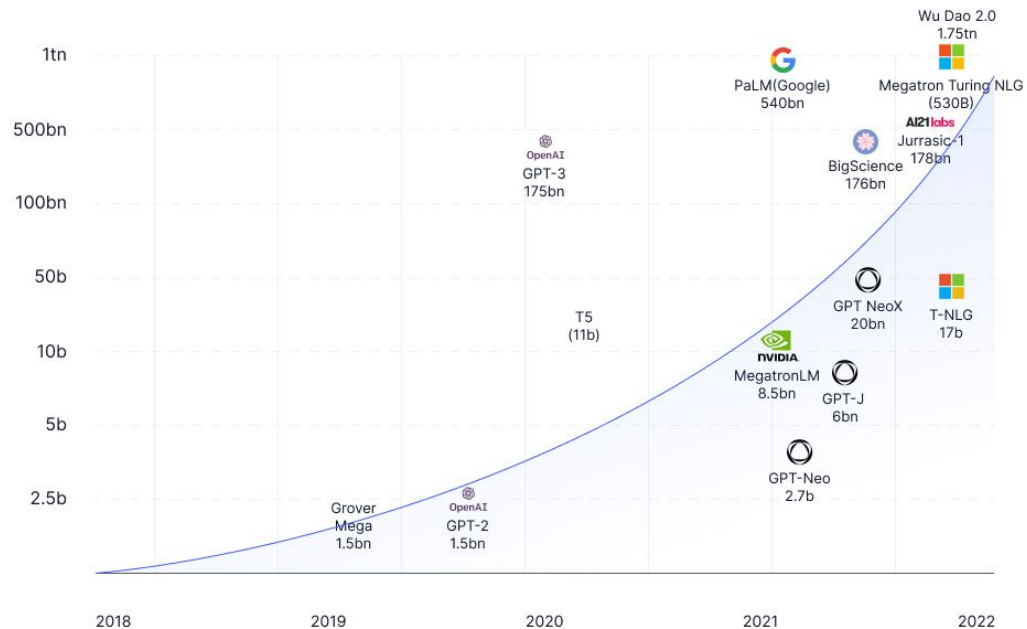


They process unstructured text as sequence of tokens



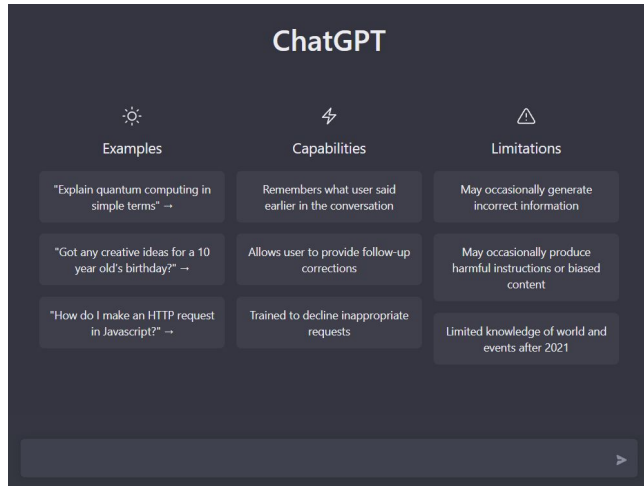
They are pretrained on exponentially growing model sizes

text.cortex



<https://textcortex.com/post/how-gpt-3-writing-tools-work>

Pretrained LLMs can answer questions, reason and generate fluent text



ChatGPT passes MBA exam given by a Wharton professor

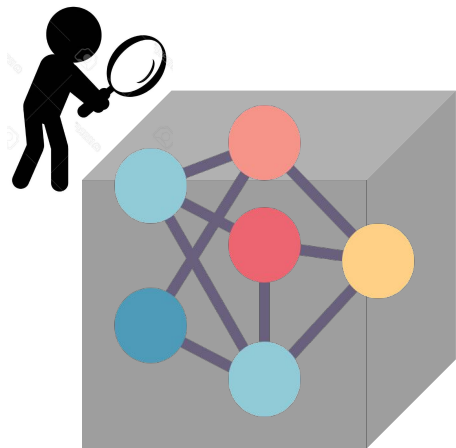
Alarmed by A.I. Chatbots, Universities Start Revamping How They Teach

Meet Bard, Google's Answer to ChatGPT

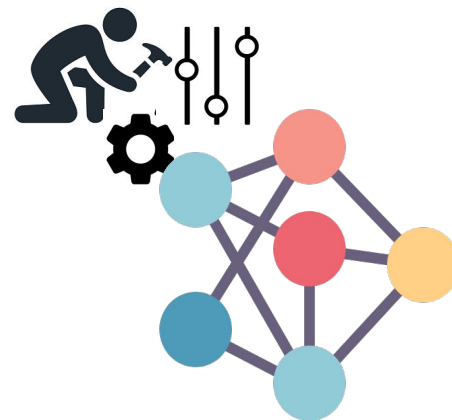
Scores of Stanford students used ChatGPT on final exams, survey suggests

ChatGPT listed as author on research papers: many scientists disapprove

Design Flaws - No transparency or control

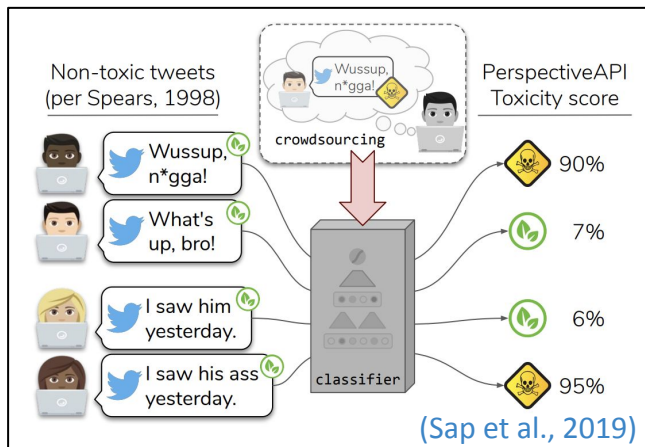


Models not transparent by design
(Lipton, 2018; Vellido, 2020; Belinkov et al., 2020)

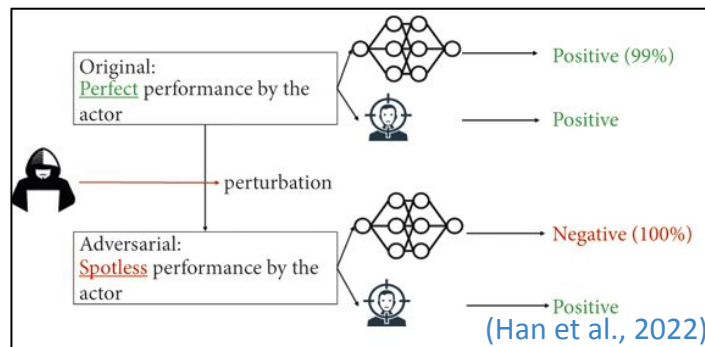


Models hard to control by design
(Ziegler et al., 2019; Dathathri et al., 2020)

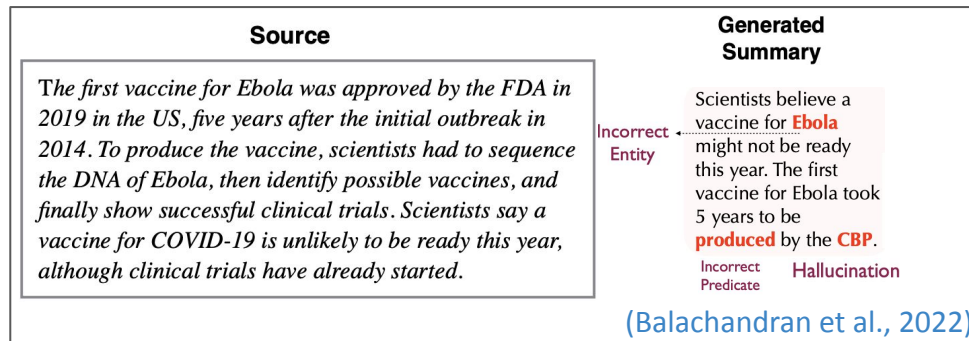
Unintended effects due to such design flaws



Spurious correlations



Low generalizability





Factually Unreliable

My Research - Trustworthy NLP Systems

Designing large-scale, data-driven language generation models to be more **trustworthy, transparent and reliable** through **interpretable and controllable data and model designs**



Factual Unreliability of Model Generated Text

 **Gael Breton** 
@GaelBreton

6/ It makes a TON of stuff up.

I decided to lightly fact check the response it gave to a journalist about picking a new phone.

The screenshot speaks for itself. I wouldn't want this level of accuracy to make buying decisions.



Hello, this is Bing. I can help you find a phone with a good camera and battery life under \$1500 AUD. 😊

According to the web, some of the best camera phones in 2023 are:

- **Google Pixel 7 Pro** ¹: It has a 50MP main camera, a 48MP telephoto camera with 4x optical zoom, and a 12MP ultra-wide camera. It also has a 12MP selfie camera and can record 8K video at 30fps. It has a 6.7-inch OLED display and a 5000mAh battery. It costs \$499 USD ² which is about \$704 AUD. **Wrong price, that's the Pixel 7**
- **Samsung Galaxy S22 Ultra** ³: It has a 108MP main camera, a 12MP telephoto camera with 10x optical zoom, and a 12MP ultra-wide camera. It also has a 40MP selfie camera and can record 8K video at 24fps. It has a 6.8-inch AMOLED display and a 5000mAh battery. It costs \$999 USD ⁴ which is about \$1408 AUD. **Wrong, it's \$1799**
- **Apple iPhone 14 Pro** ⁴: It has a 12MP main camera, a 12MP telephoto camera with 3x optical zoom, and a 12MP ultra-wide camera. It also has a 12MP selfie camera and can record 8K video at 30fps. It has a 6.1-inch OLED display and a 3095mAh battery. It costs \$999 USD ⁴ which is about \$1408 AUD. **Wrong, it's 48MP**

These are some of the options you can consider. You can also check out the shopping results for more choices ².

3:20 AM · Feb 8, 2023 · 180.9K Views

 **François Chollet** 
@fchollet

People in tech characterize this as "LLMs make factual errors", but that's a misleading framing, implying that LLMs have a model of what they say and this model is sometimes wrong.

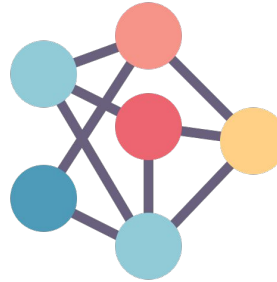
For a LLM there is no difference between saying something true, something false, or pure nonsense.

ARTIFICIAL INTELLIGENCE / TECH / GOOGLE

Google's AI chatbot Bard makes factual error in first demo

/ The mistake highlights the biggest problem of using AI chatbots to replace search engines – they make stuff up.

Automatic Text Summarization



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 instead report information collected by the C.D.C. on its virus tracking pages.

Summary

Automatic Text Summarization



The New York....
communities.

Source Document



As local data sources
pages.

Summary



The New York Times



Elicit

The Times Switches to C.D.C. Covid Data, Ending Daily Collection

As local data sources become less reliable, The Times will instead report information collected by the C.D.C. on its virus tracking pages.

After more than three years of daily reporting on the number of Covid-19 cases and deaths in every county in the United States, The New York Times is ending its Covid data-gathering operation. The Times will continue to publish its [Covid tracking pages](#) for the United States, only now they will be based on the latest information available from the [federal government](#), not the Times's data set.

The tracking pages will still show data about hospital patients with Covid; reported cases and tests; and how many people have died from the virus. Data on vaccination rates and comparisons between vaccinated and unvaccinated populations will also remain.

A new interactive county map will show local levels of

What is the impact
of creatine on
cognition?

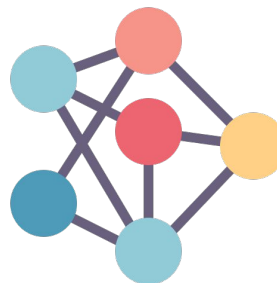
beta



SUMMARY OF TOP 4 PAPERS

These papers have mixed findings on creatine's impact on cognition. Three of these papers suggest that creatine may improve cognition: [Ling 2009](#) found that creatine supplementation may improve performance on some cognitive tasks, [McMorris 2007](#) found that creatine supplementation aids cognition in the elderly, and [Benton 2010](#) found that in vegetarians, creatine supplementation resulted in better

Factual Accuracy of Model Generated Summaries



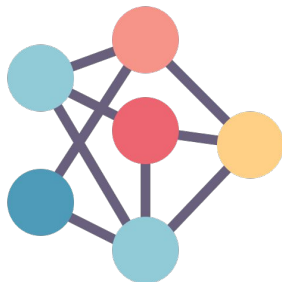
The New York Times is ending its Covid data-gathering operation. The Times will continue to publish its Covid tracking impact of the virus on communities.

Source Document

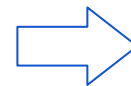
As local data sources become less reliable, The Times will **stop reporting** information collected by the C.D.C. on its **pandemic headlines**.

**Summary w/
Factual Errors**

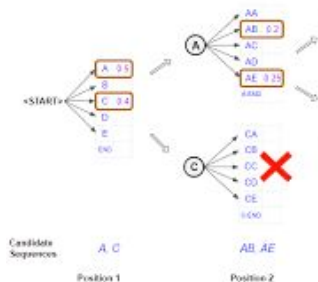
Interventions to Mitigate Factual Errors



Factuality Constraints and Objectives (Cao et al, 21)



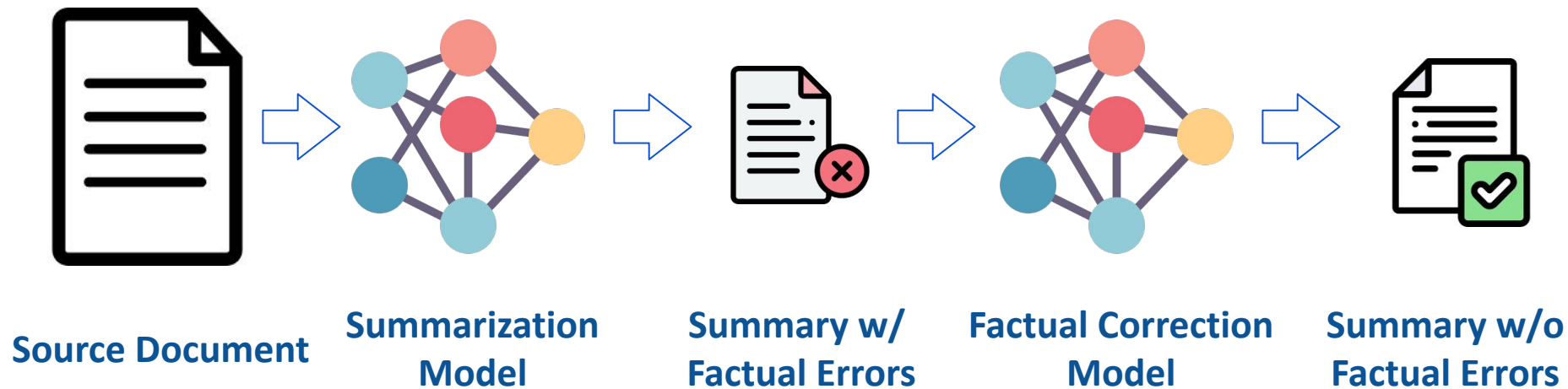
Summary w/ & w/o Factual Errors



Decoding Constraints (King et al, 22)

Source Document

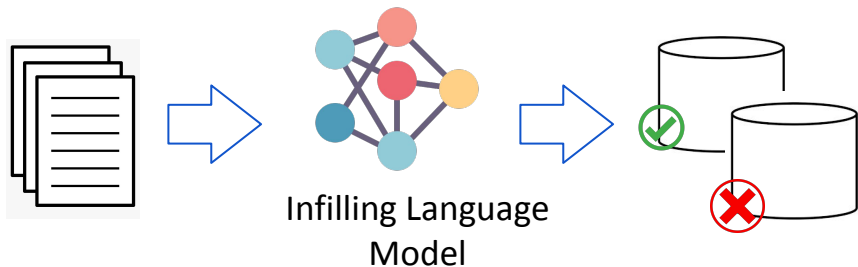
Post-Editing to Correct Factual Errors



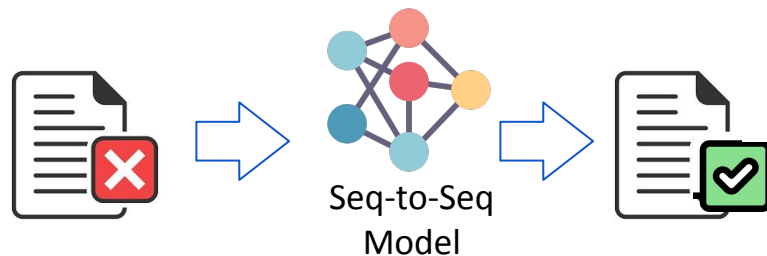
Challenges in Collecting Training Data

- Training Data: (Incorrect Summary, Correct Summary) Pairs
- Human Annotated Data
 - Expensive - Long Process to read and edit summaries
 - Subjective - Factuality is subjective
- Synthetic Data - Create adversarial data (incorrect summaries) using heuristic rules ([Kryściński, et. al, 2020](#), [Cao, et. al, 2020](#))
 - Low Generalizability - Only able to correct few heuristic errors trained on
 - Low Coverage - Hard to design heuristics for complex factual errors

FactEdit - Infilling LMs for Syntactic Adversarial Data Generation



Infilling based Generation of Adversarial Summaries



Factual Error Correction

Editing at Fine-Grained Unit Level

[Vaccine for Covid-19]_{Subj} [is]_{Rel}
[unlikely to be ready this
year.]_{Obj}

[Vaccine for Ebola]_{Subj} [is]_{Rel}
[unlikely to be ready this
year.]_{Obj}

[Vaccine for Covid-19]_{Subj} [is]_{Rel}
[under clinical trials.]_{Obj}

[Vaccine for Covid-19]_{Subj} [looks less]_{Rel}
[unlikely to be ready this year.]_{Obj}

Infilling LMs for Candidate Generation

Reference: Vaccine for Covid-19 is unlikely to be ready this year.

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

The first vaccine for [MASK] is
[SEP] clinical trials ...



Covid-19

Infilling Language
Model

Infilling LMs for Candidate Generation

Reference: Vaccine for Covid-19 is unlikely to be ready this year.

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

Vaccine for [MASK] is unlikely to be ready this year, started. [SEP] For reference the vaccine ...



Infilling Language Model

Coronavirus ✓

Covid-19 ✓

Ebola ✗

Polio ✗

Vaccine for Ebola is unlikely to be ready this year.

Vaccine for Polio is unlikely to be ready this year.

Beam-Search Candidates

Infilling LM - Training Data Generation

Incorrect: Vaccine for Ebola is unlikely to be ready this year.

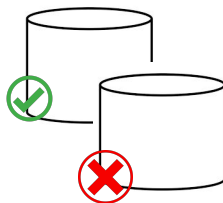
Incorrect: Vaccine for Covid-19 is being prepared for this year.

Incorrect: Researchers expect to begin Phase 3 trials after production.

Correct: Vaccine for Covid-19 is unlikely to be ready this year.

Correct: Vaccine for Covid-19 is unlikely to be ready this year.

Correct: Researchers expect to begin production after Phase 3 trials.

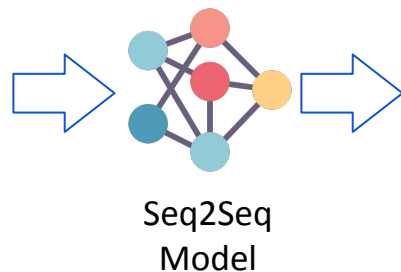


Training Phrase-Level Error Correction

Summary: Vaccine for Ebola is unlikely to be ready this year.

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



Vaccine for Covid19 is unlikely to be ready this year.

Phrase-Level Error Correction

Incorrect Model Generated

Summary: World Leaders met Ban Ki Moon for UN Meeting in 2020.

Context: UN Sec. Gen. Antonio Gutteres met the leaders for economic implications of global sanctions.

World Leaders met Ban Ki Moon for UN Meeting in 2020 [SEP]

Pandemic Response ... [SEP] UN Sec. Gen. Antonio Gutteres met the leaders for ...



Seq2Seq
Model

World Leaders met Antonio Gutteres for UN Meeting in 2020.

Data & Experiment Settings

- Standard Datasets - CNN/DM + XSum
 - Full Test Set evaluation using BART Model
- FRANK Benchmark - human annotated judgements for error categorization
 - Generated outputs from **5 models for CNN/DM**
 - Generated outputs from **4 models for XSum**
- Evaluation
 - Rouge - fluency, grammar and content
 - FactCC - semantic, sentence level errors
 - Ent-DAE - discourse errors

Post-Editing factual errors improves factuality of generated summaries

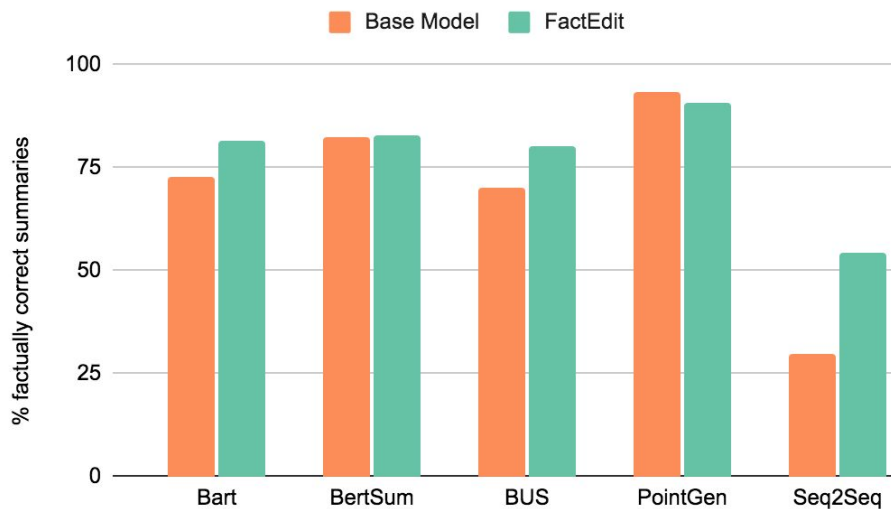
Dataset	Method	R1	R2	RL	FactCC	Ent-DAE
CNN/DM	Bart (Lewis et al., 2020)	44.07	21.08	41.01	75.78	74.85
	Cao (Cao et al., 2020)	42.72	20.59	39.92	49.98	74.83
	FACTEDIT	42.17	20.22	39.37	75.49	75.71
	FACTEDIT + FactCC Filter (FF)	42.53	20.48	39.74	76.03	75.36

Post-Editing factual errors improves factuality of generated summaries

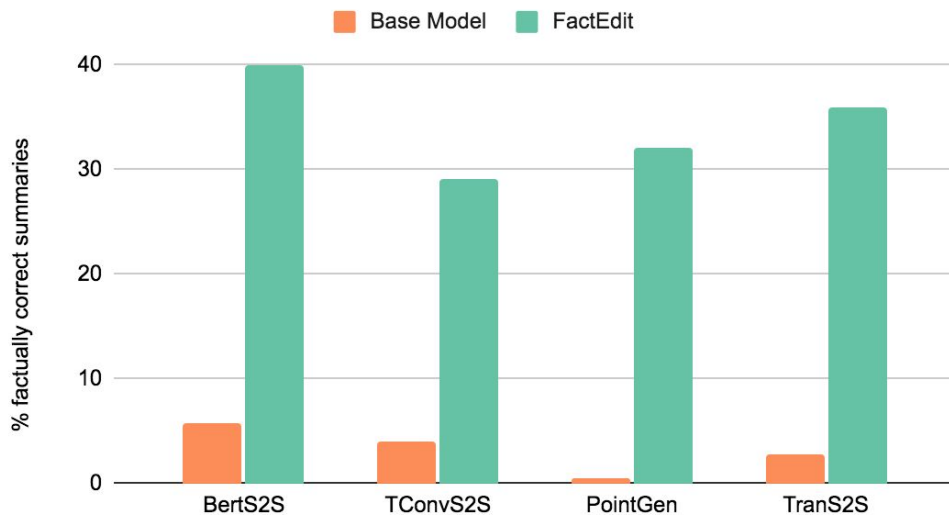
Dataset	Method	R1	R2	RL	FactCC	Ent-DAE
CNN/DM	Bart (Lewis et al., 2020)	44.07	21.08	41.01	75.78	74.85
	Cao (Cao et al., 2020)	42.72	20.59	39.92	49.98	74.83
	FACTEDIT	42.17	20.22	39.37	75.49	75.71
	FACTEDIT + FactCC Filter (FF)	42.53	20.48	39.74	76.03	75.36
XSum	Bart (Lewis et al., 2020)	34.71	15.04	27.40	21.93	20.03
	Cao (Cao et al., 2020)	33.64	14.71	26.49	7.01	20.03
	FACTEDIT	33.58	14.68	26.71	23.91	20.13
	FACTEDIT + FactCC Filter (FF)	33.58	14.68	26.71	23.91	20.13

Infilling based generated data generalizes correction across models

Performance across Models on CNN/DM



Performance across Models on XSum



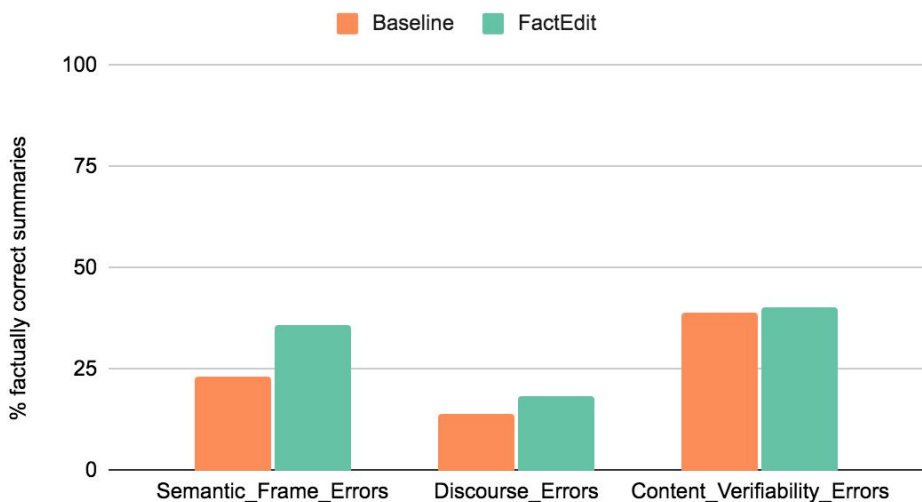
FactEdit improves performance across most data and model settings
Large Gains in abstractive XSum Dataset

Infilling based generated data generalizes correction across error types

Performance across Error Types on CNN/DM



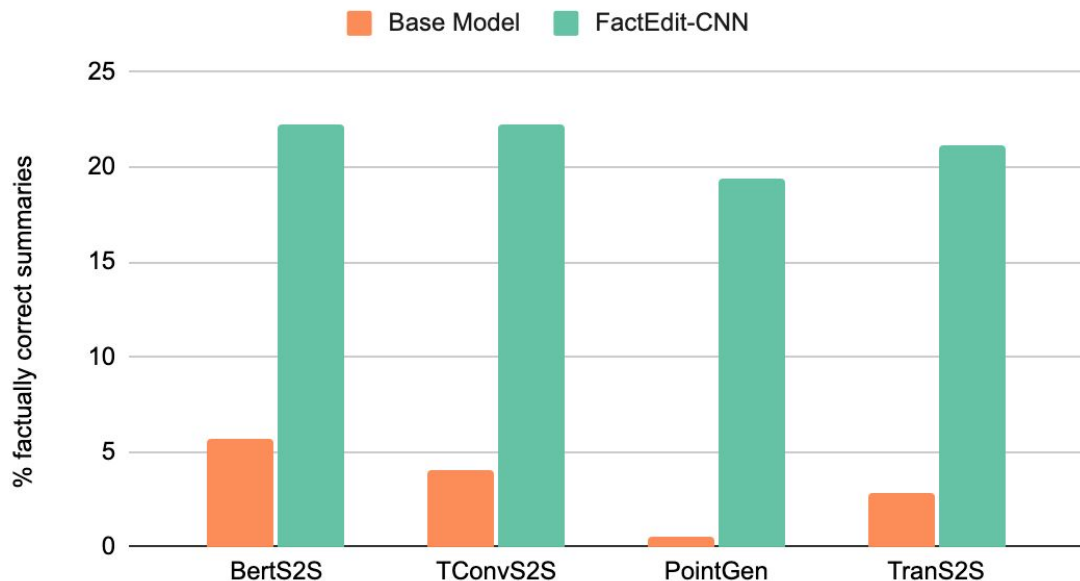
Performance across Error Types on XSum



FactEdit improves performance across error types
Especially good at correcting semantic frames.

Infilling based generated data generalizes correction across datasets

Performance across Models on XSum



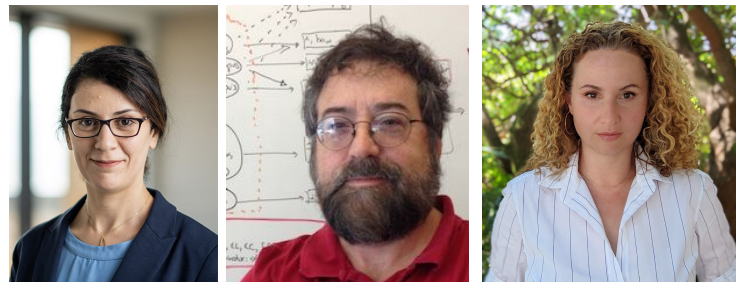
FactEdit trained on data from longer summaries (CNN/DM) improves results on headline generation (XSum)

Summary

- **Data Generation leveraging Infilling LMs to edit fine-grained phrases**
 - Editing at syntactic units provides control over data generation process
 - Infilling LMs provide challenging, generalizable generation process
- **FactEdit - Phrase-level Factual Error Correction**
 - Syntactic Phrase based training data enables fine-grained correction
 - Correct Factual Errors without modifying other generated text
- **Generalizable** across models and error categories
 - Generated training data includes diverse examples of errors
 - Flexible and Diverse data generation generalizes error correction across multiple models and error categories

Thank you!

- Data Generation leveraging **Infilling LMs to edit fine-grained syntactic units**
 - Editing at syntactic units provides control over data generation process
 - Infilling LMs provide challenging, generalizable generation process
- FactEdit - **Phrase-level Factual Error Correction**
 - Syntactic Phrase based training data enables fine-grained correction
 - Correct Factual Errors without modifying other generated text
- **Generalizable** across models and error categories
 - Generated training data includes diverse examples of errors
 - Flexible and Diverse data generation generalizes error correction across multiple models and error categories



Questions: vbalacha@cs.cmu.edu



Paper

Code