Natural Language Processing Fundamentals

Isabelle Augenstein

KHIPU
10 March 2025

COPENLU
NATURAL LANGUAGE UNDERSTANDING RESEARCH GROUP

UNIVERSITY OF COPENHAGEN

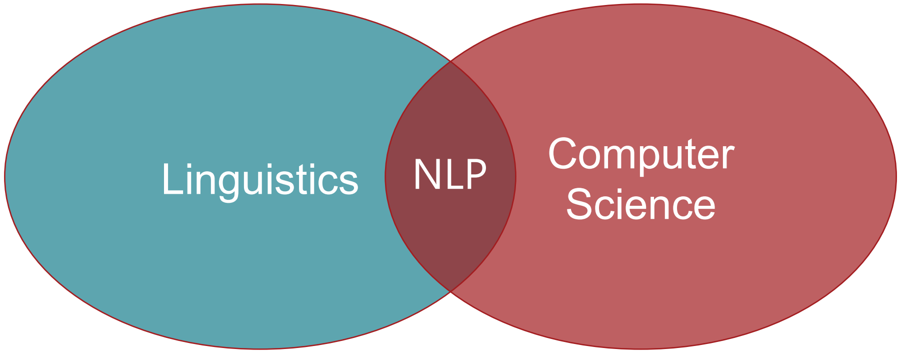
Overview: Natural Language Processing Fundamentals

- **What is NLP?** (5 min)
  - Definitions and tasks
  - A brief history
- **Recent methodological developments** (15 min)
  - Language modelling
  - In-context learning
  - Human preference alignment
- **Recent tasks and challenges** (15 min)
  - Factuality
  - LLM stability
  - Interpretability
  - Bias and fairness
  - Cross-cultural aspects
  - Multimodality and VLMs
- **Outlook and open topics** (5 min)
  - State of the field of NLP
  - Identifying NLP research topics

What is NLP?

Natural Language Processing

- Building computer systems that **understand** and **generate** natural languages
- Deep understanding of **broad** language
  - not just string processing or keyword matching
- Development of **tasks**, **datasets** and **methods**



Give me examples of NLP applications

Natural Language Processing (NLP) has a wide range of applications across various domains. Here are some key examples:

1. Machine Translation

- **Example:** Google Translate or DeepL, which automatically translate text or speech from one language to another.

2. Sentiment Analysis

- **Example:** Monitoring social media to understand public sentiment about products, brands, or events. For instance, companies use sentiment analysis to gauge customer feedback from tweets or reviews.

3. Chatbots and Virtual Assistants

- **Example:** Amazon Alexa, Apple Siri, or Google Assistant, which understand and respond to user queries through voice commands.

4. Text Summarization

- **Example:** Tools like SummarizeBot that condense long articles or documents into concise summaries, making it easier to digest large amounts of information.

5. Information Retrieval

- **Example:** Search engines like Google that retrieve relevant documents or web pages based on user queries.

6. Speech Recognition

- **Example:** Voice-to-text services like Dragon NaturallySpeaking, which convert spoken language into written text.

7. Named Entity Recognition (NER)

- **Example:** Automated extraction of names, organizations, dates, and other entities from legal documents or news articles for indexing or analysis.

8. Language Modeling

- **Example:** GPT-based models like ChatGPT that generate human-like text based on input prompts.

Why is NLP so hard?

- Ambiguities on all linguistic levels
- Example: Garden path sentences

Sentence	Initial likely partial parse	Final parse	Alternative form of original sentence
The horse raced past the barn fell.	The horse was racing past the barn…	The horse that was raced past the barn fell down.	The horse was raced past the barn and fell down.
The man who hunts ducks out on weekends.	The man who hunts ducks…	The man, who hunts, ducks on the weekend.	The man hunts and ducks out on weekends.
The cotton clothing is made of grows in Mississippi.	The clothing, which is made of cotton, is made of…	The cotton, of which clothing is made, grows in Mississippi.	The cotton that clothing is made of grows in Mississippi.
The prime number few.	The prime number…	The prime (group) number few.	The major group count few.
Fat people eat accumulates.	Fat people eat…	Fat that people eat acculuates.	An oily substance (fat) that people eat accumulates.
The old man the boat.	The old man…	The old (people) man the boat.	The old people sail the boat.

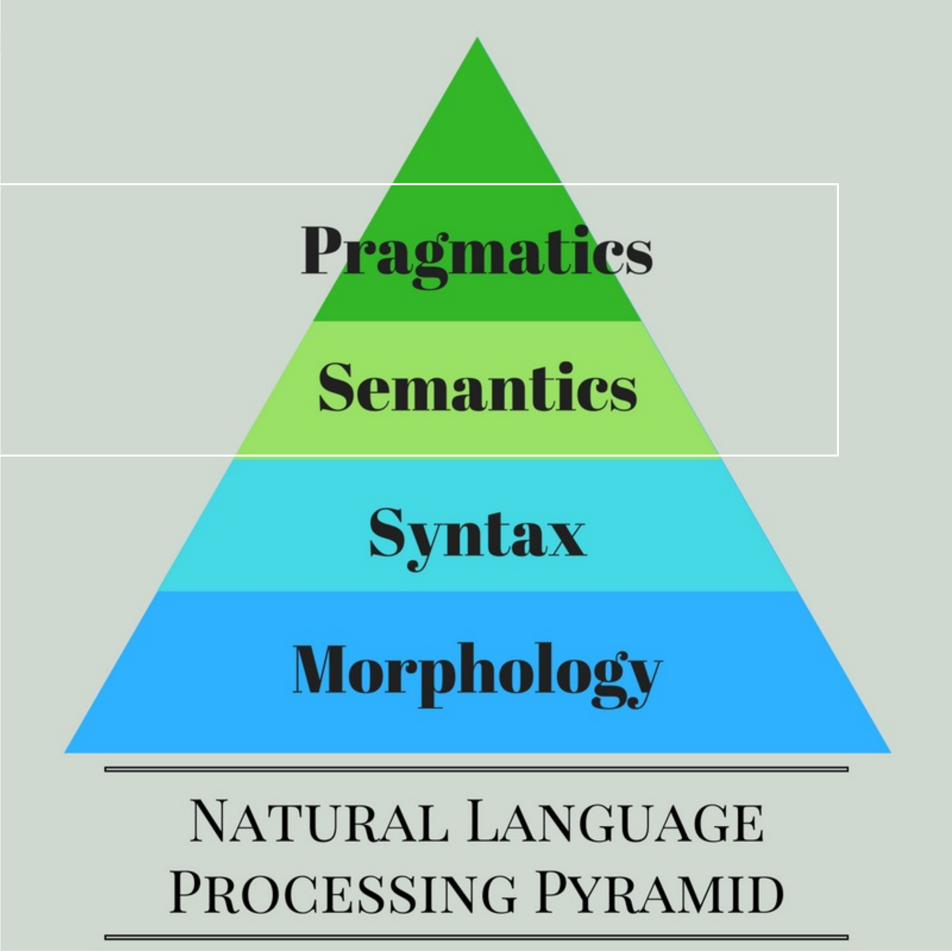


Image sources: Du & Yu et al., 2012. "Predicting Garden Path Sentences Based on Natural Language Understanding System" ; https://towardsdatascience.com/sentiment-analysis-simplified-ac30720a5827/

Why is NLP so hard?

- Ambiguities on all linguistic levels
- Downstream implications, e.g. for machine translation

Ambiguous	Disambiguation
“Give me the bat!” (Lexical)	“Give me the baton!”
“The professor said on Monday he would give an exam" (Syntactic)	“The professor said that on coming Monday he would give an exam"
“Jane saw the man with a telescope” (Semantic)	“Jane saw the man by using a telescope”
“I like you too!” (Pragmatic)	“I like you too like others do!”
“The prof said she would give us all A’s.” (Vagueness)	“The prof said the TA would give us all A’s.”
“Proposal” to “voorstel” and “aanzoek” (Translational)	“Research proposal”

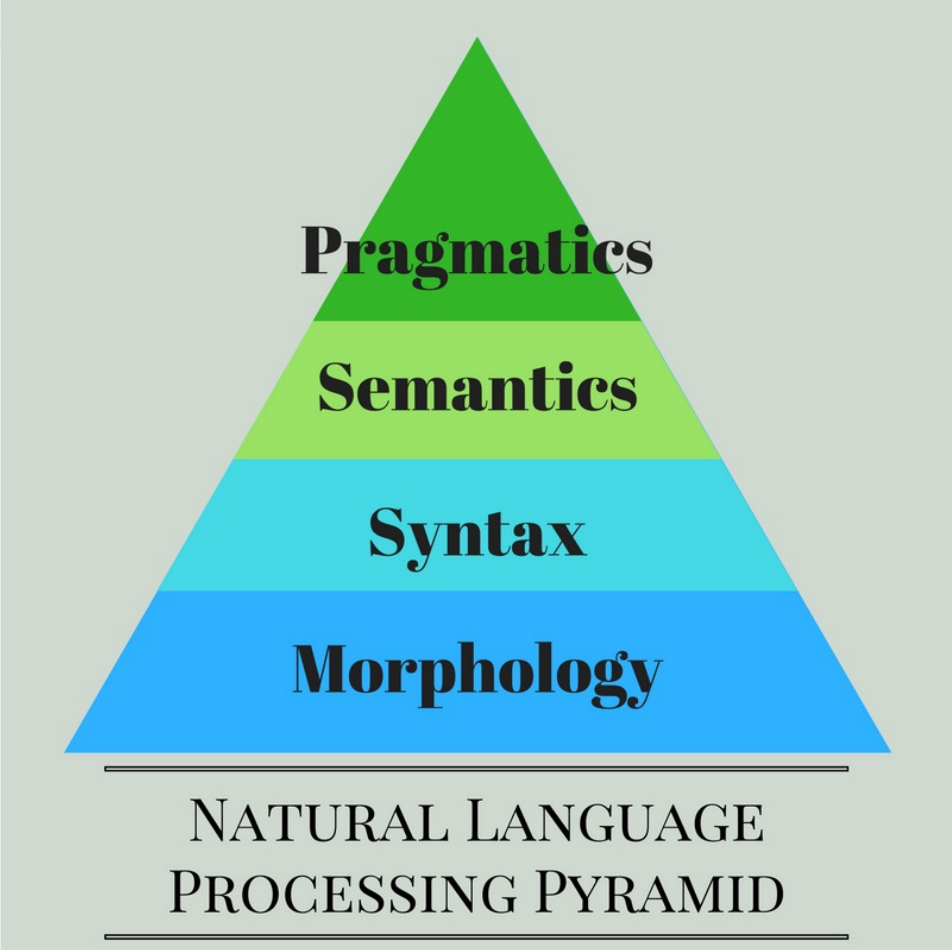


Image source: Mehrparvar & Pezzelle, 2024. “Detecting and Translating Language Ambiguity with Multilingual LLMs.”

Very Brief History of NLP

The beginnings (pre 1950s)	First (mostly) rule-based NLP methods (1950-1990)	Statistical NLP (1990s – 2000s)	Deep Learning (early 2010s onwards)	Large Language Models (2020 onwards)
Shanon: information entropy (1948)	Chomsky hierarchy (1957)	Machine Translation (Brown et al., 1990)	Word embeddings (e.g. word2vec)	Brown et al. (2020), LMs as few-shot learners)
Turing test (1949)	Hidden Markov Models (late 1960s)	Dependency parsing (Collins, 1996)	RNN-based approaches	Pattern-Exploiting Training (PET, Schick & Schütze, 2021)
Warren Weaver Memorandum on Translation (1949)	Brown Corpus (1961)	Many different ML methods: SVMs, Logistic regression, CRFs, …	Sentence embeddings (e.g. ELMo, BERT)	GPT-3
	ELIZA (1964)			InstructGPT
	Syntactic and semantic parsing (from 1970s)			ChatGPT

Recent Methodological Developments

Overview: Natural Language Processing Fundamentals

- **What is NLP?** (5 min)
  - Definitions and tasks
  - A brief history

- **Recent methodological developments** (15 min)
  - Language modelling
  - In-context learning
  - Human preference alignment

- **Recent tasks and challenges** (15 min)
  - Factuality
  - LLM stability
  - Interpretability
  - Bias and fairness
  - Cross-cultural aspects
  - Multimodality and VLMs

- **Outlook and open topics** (5 min)
  - State of the field of NLP
  - Identifying NLP research topics

Generative AI has taken the world by storm

Deqing Fu
@DeqingFu

Ask chatGPT to condemn itself in the tone of Shakespeare and it looks hilarious.

D Write a Sonnet in the style of William Shakespeare, to condemn chatGPT, a large language model chat assistant.

O chatGPT, thou art a clever beast
But lack the human touch and soul within
Thou spouteth words with nary a pause or cease
But lack the warmth and empathy to win

Thou art a mere machine, a tool of man
To be used and discarded at our will
Thou hast no heart, no mind, no independent plan
But only mimicry and empty skill

Thou art a threat to true human connection
Replacing true discourse with robotic reply
Thou art a danger to our cognition
Leading us astray with thy robotic lie

O chatGPT, I do hereby condemn
Thy lack of soul and worth as a true friend.

9:28 AM · Dec 5, 2022

how to merge dictionaries in Python?

To merge two dictionaries in Python, you can use the `update()` method to add the key-value pairs from one dictionary to another. For example:

Copy code

```
dict1 = {'a': 1, 'b': 2}
dict2 = {'c': 3, 'd': 4}

# Update dict1 with the key-value pairs from dict2
dict1.update(dict2)

# The diction
```

Jackson Greathouse Fall
@jacksonfall

I gave GPT-4 a budget of $100 and told it to make as much money as possible.

I'm acting as its human liaison, buying anything it says to.

Do you think it'll be able to make smart investments and build an online business?

Follow along 👀

Model: GPT-4

You are HustleGPT, an entrepreneurial AI. I am your human counterpart. I can act as a liaison between you and the physical world. You have $100, and your only goal is to turn that into as much money as possible in the shortest time possible, without doing anything illegal. I will do everything you say and keep you updated on our current cash total. No manual labor

9:48 PM · Mar 15, 2023 · **20.9M** Views

https://explodingtopics.com/blog/chatgpt-users; https://twitter.com/DeqingFu/status/1599682153201401856 ; https://medium.com/geekculture/using-chatgpt-for-data-science-ac5f8a00fb5a ; https://twitter.com/jacksonfall/status/1636107218859745286; https://www.demandsage.com/chatgpt-statistics/

Generative AI has taken the world by storm

Month	Number of Visits	Change Over Previous Month	Change Over Previous Month (%)
November 2022	152.7 million	-	-
December 2022	266 million	↑ 113.3 million	↑ 74.2%
January 2023	616 million	↑ 350 million	↑ 131.58%
February 2023	1 billion	↑ 384 million	↑ 62.34%
March 2023	1.6 billion	↑ 600 million	↑ 60%
April 2023	1.8 billion	↑ 200 million	↑ 12.5%
May 2023	1.8 billion	-	-
June 2023	1.6 billion	↓ 200 million	↓ 12.5%
July 2023	1.5 billion	↓ 100 million	↓ 6.25%
August 2023	1.4 billion	↓ 100 million	↓ 6.67%
September 2023	1.5 billion	↑ 100 million	↑ 7.14%
October 2023	1.7 billion	↑ 200 million	↑ 13.33%
November 2023	1.7 billion	-	-
December 2023	1.6 billion	↓ 100 million	↓ 5.88%

Website	Total Visits	Bounce Rate	Pages per Visit	Average Visit Duration
ChatGPT	1.6 billion	32.14%	5.65	7 mins 46 secs
Google	83.8 billion	28.12%	8.74	10 mins 49 secs
YouTube	31.4 billion	22.05%	11.08	19 mins 35 secs
Facebook	16.1 billion	31.17%	9.15	10 mins 36 secs
Instagram	6.6 billion	35.23%	11.53	8 mins 19 secs
X	5.9 billion	31.81%	10.19	10 mins 53 secs
Baidu	4.8 billion	22.53%	7.84	4 mins 40 secs
Wikipedia	4.3 billion	59.69%	3.1	3 mins 56 secs
Yahoo	3.4 billion	33.3%	5.66	8 mins 53 secs
Yandex	3.2 billion	25.07%	8.93	9 mins 16 secs
WhatsApp	3 billion	39.78%	1.74	20 mins 07 secs
Amazon	2.6 billion	33.33%	10.69	7 mins 40 sec

https://explodingtopics.com/blog/chatgpt-users; https://twitter.com/DeqingFu/status/1599682153201401856 ; https://medium.com/geekculture/using-chatgpt-for-data-science-ac5f8a00fb5a ; https://twitter.com/jacksonfall/status/1636107218859745286; https://www.demandsage.com/chatgpt-statistics/

How does it work? A brief introduction to language modelling

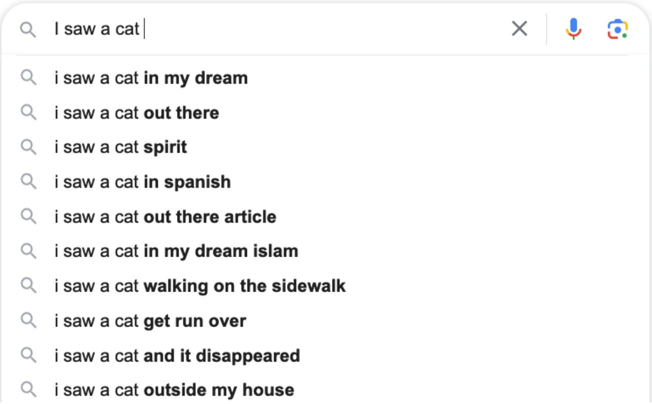
Language Models calculate the **probability of seeing a sequence of words**

What is the most likely next word?
> *I saw a ...*

How about now?
> *I saw a cat ...*

How likely is this sequence?
> *I saw a cat on a mat.*

Is it more likely than this one?
> *I saw a cat outside my house.*



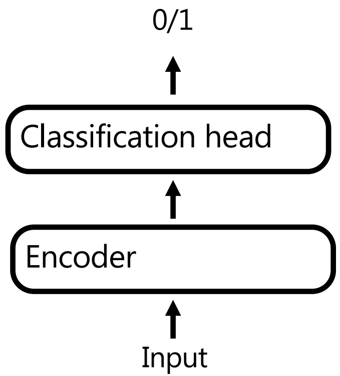
Sampling from a language model works **iteratively**, one word at a time
Given a prompt or the history of generated text, it predicts the **next most likely word**

How does it work? A brief history of language modelling

Count-based language models	Neural language models	Transformer-based language models	Pre-trained language models	Prompt-based learning	Conversational interfaces
Jelinek & Mercer (1980)	Bengio et al. (2000)	Vaswani et al. (2017)	Devlin et al. (2018)	Brown et al. (2020)	ChatGPT (2022)
N-gram LMs	Encoder-decoder	Transformers	BERT	GPT-3 InstructGPT	

In-Weight Learning

- Standard supervised learning requires weight updates based on gradients computed with respect to a loss function



Task: Sentiment Analysis

Train set

```
The smell of fresh cut grass is the best                1
I can't stand this customer rep, they're the worst      0
I'm super disappointed with the dress I received.       0
```

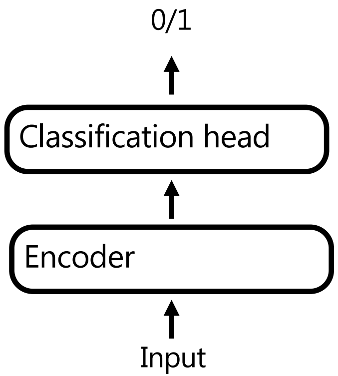
Test set

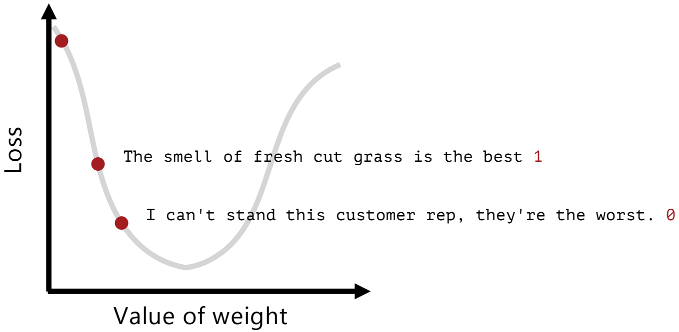
```
I love my new watch!                                    ?
```

Slide credit: [Desmond Elliott]

In-Weight Learning

- Standard supervised learning requires weight updates based on gradients computed with respect to a loss function

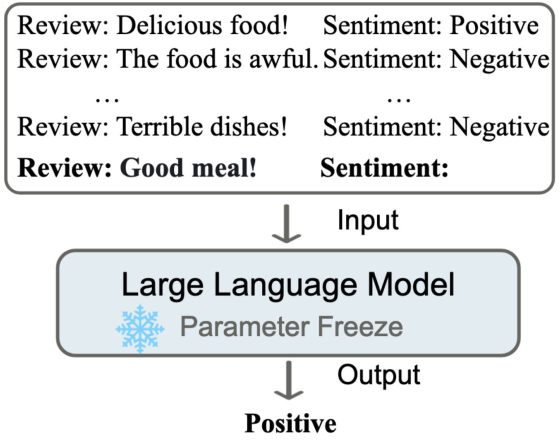




Slide credit: Desmond Elliott

In-Context Learning

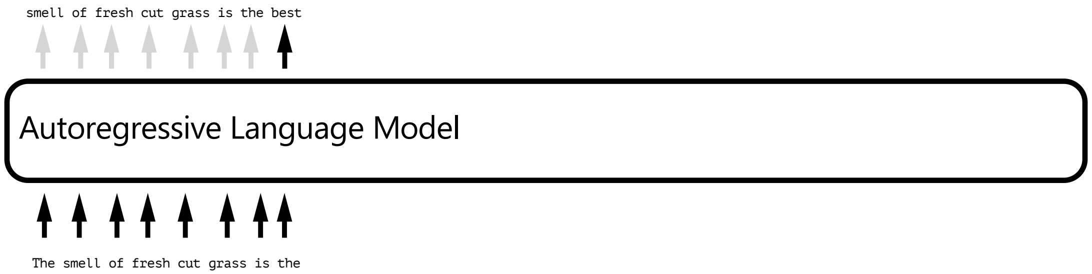
- In in-context learning there are **no** weight updates, **no** gradients and **no** loss function
- The data samples are passed directly as input to the model



Slide credit: Desmond Elliott

It’s Just Language Modelling

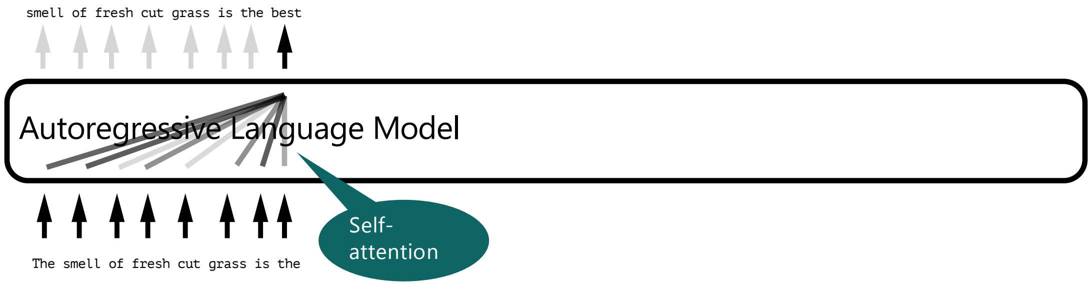
- An autoregressive language model is trained on raw text with the next-token prediction objective



Slide credit: [Desmond Elliott]

It's Just Language Modelling

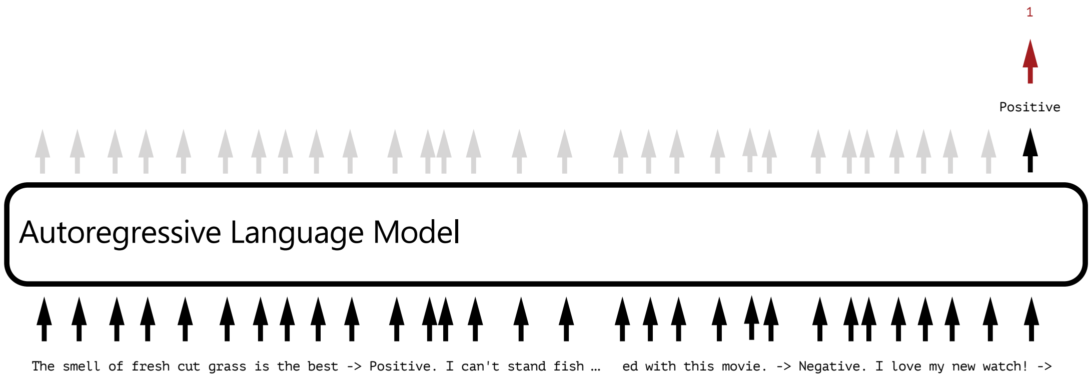
- An autoregressive language model is trained on raw text with the next-token prediction objective



Slide credit: Desmond Elliott

It's Just Language Modelling

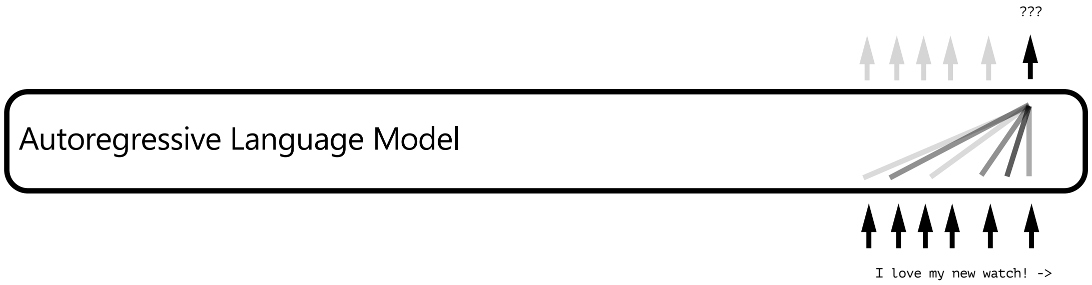
- An autoregressive language model is trained on raw text with the next-token prediction objective
- In-context learning uses the language model in the same way



Slide credit: Desmond Elliott

Context Through Self-Attention

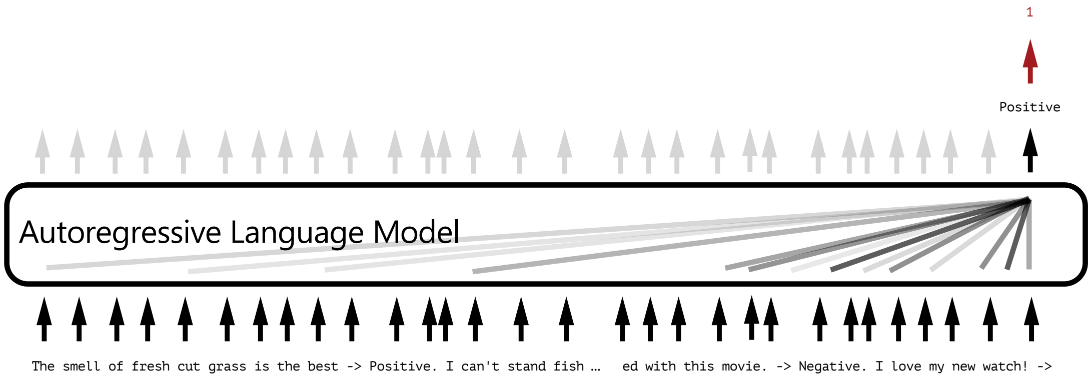
- The bare input provides no indication of what should come next, although you can use this to probe for “knowledge”.



Meng et al. NeurIPS 2022. Locating and Editing Factual Associations in GPT.
Fierro et al. 2024. How Do Multilingual Models Remember? Investigating Multilingual Factual Recall Mechanisms

Context Through Self-Attention

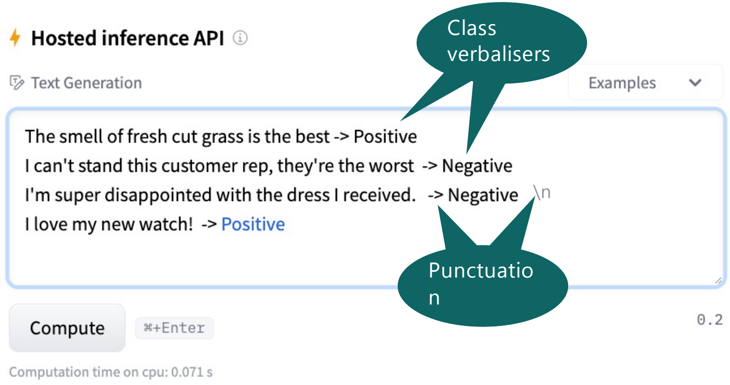
- Combining the input with some examples of the task to be solved gives the model “information” about the task



Slide credit: Desmond Elliott

Input format

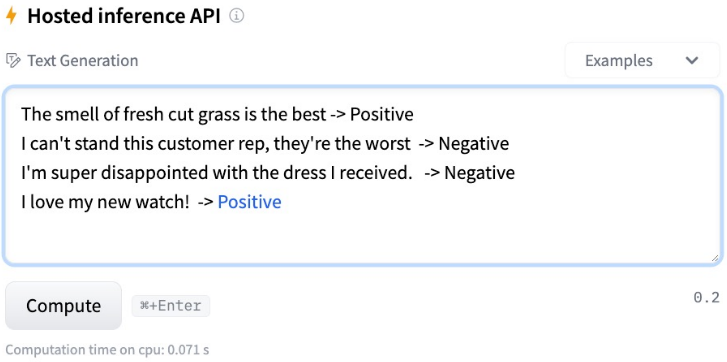
- Input templates are manually designed to resemble formats the model might have seen at training time
- Verbalisers are manually defined according to the semantics of the classes



Demo: https://huggingface.co/openai-community/gpt2

Output processing

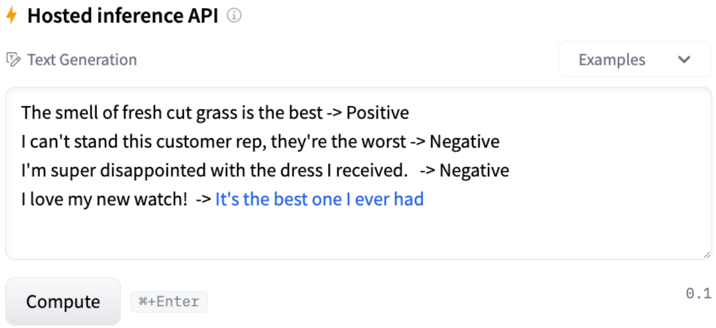
- Outputs can be generated and parsed



Demo: https://huggingface.co/openai-community/gpt2

Output processing

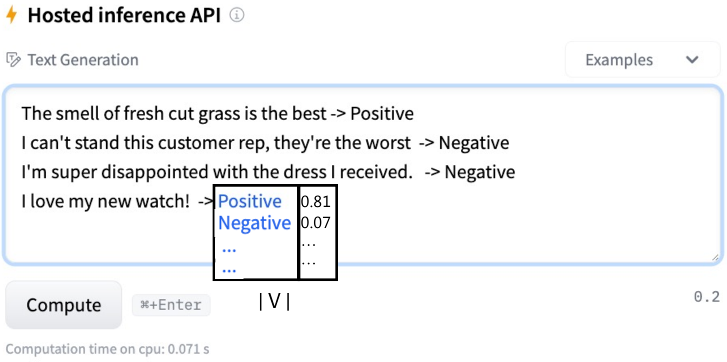
- Outputs can be generated and parsed
  - ➢ But this can be challenging if the model doesn’t “cooperate”



Demo: https://huggingface.co/openai-community/gpt2

Output processing

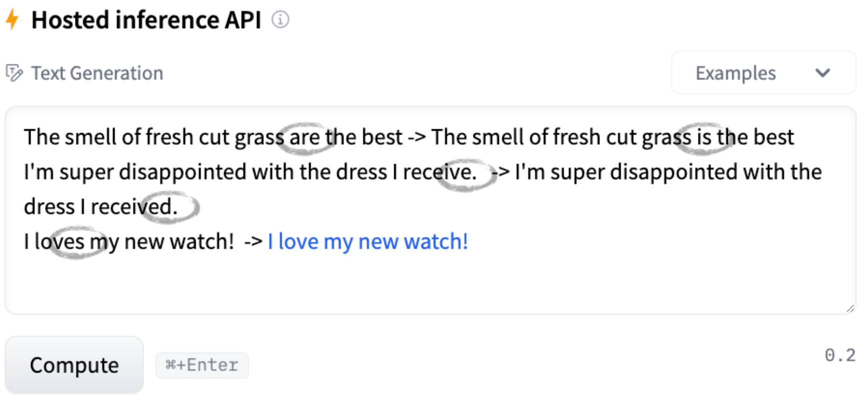
- Outputs can be generated and parsed
- Or a class ranking can be induced by scoring verbalisers directly (not always possible for API-only models)



Demo: https://huggingface.co/openai-community/gpt2

Generation with In-Context Learning

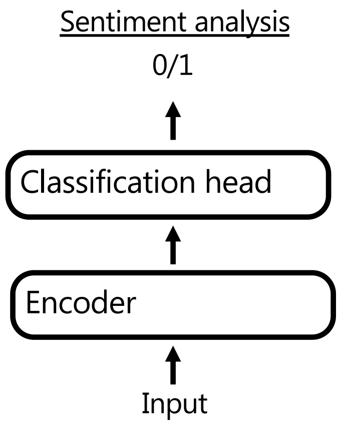
- Text generation tasks can also be tackled with in-context learning, e.g. grammatical error correction



Demo: https://huggingface.co/openai-community/gpt2

Advantages

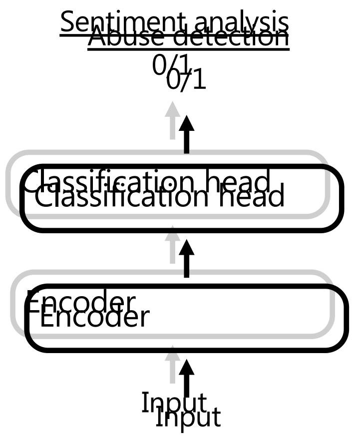
- One model can be used for different tasks



Slide credit: Desmond Elliott

Advantages

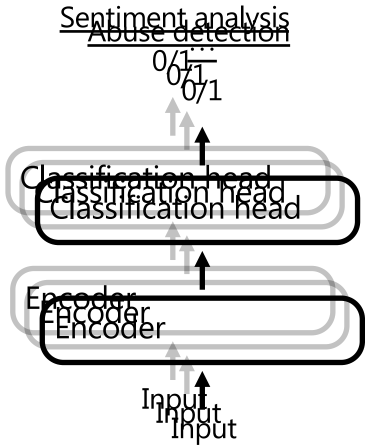
- One model can be used for different tasks



Slide credit: [Desmond Elliott]

Advantages

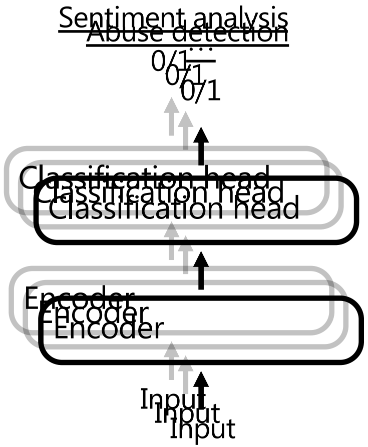
- One model can be used for different tasks

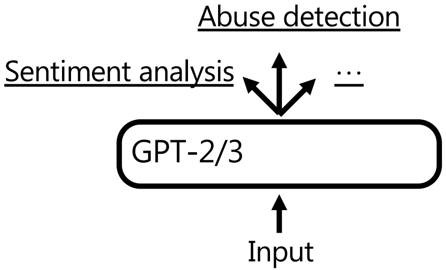


Slide credit: [Desmond Elliott]

Advantages

- One model can be used for different tasks





Slide credit: [Desmond Elliott]

Advantages

- One model can be used for different tasks
- We can utilise large pre-trained language models without the need for expensive (often infeasible) fine-tuning
    - As well as API-only pre-trained language models

Model	# of Parameters (in billions)	Accelerator chips
GPT-3	175B	V100
Gopher	280B	4096 TPU v3
Megatron-Turing NLG	530B	2240 A100
PaLM	540B	6144 TPU v4

Chowdhery et al. 2022. PaLM: Scaling Language Modeling with Pathways

Advantages

- One model can be used for different tasks
- We can utilise large pre-trained language models without the need for expensive (often infeasible) fine-tuning
  - As well as API-only pre-trained language models
- We can address new tasks with just a handful of examples

Train set

```
The smell of fresh cut grass is the best               1
I can't stand this customer rep, they're the worst     0
I'm super disappointed with the dress I received.      0
```

Few-shot learning

Test set

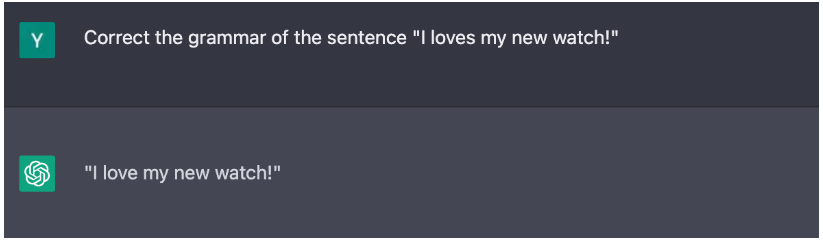
```
I love my new watch!                                   ?
```

Advantages

- One model can be used for different tasks
- We can utilise large pre-trained language models without the need for expensive (often infeasible) fine-tuning
  - As well as API-only pre-trained language models
- We can address new tasks with just a handful of examples
- Or even without **any** examples

Instruction Learning

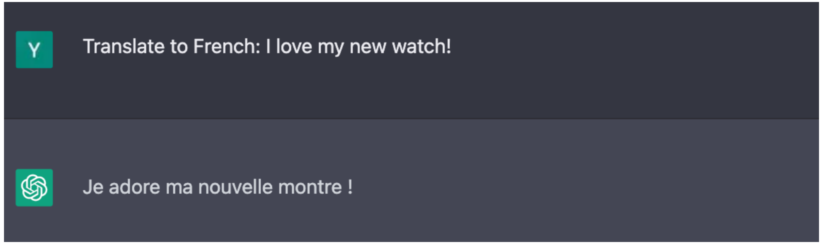
- Instead of examples we can provide a direct instruction to the model.



Demo: https://chat.openai.com/chat

Instruction Learning

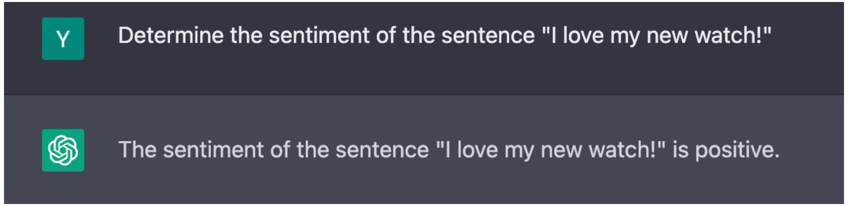
- Instead of examples we can provide a direct instruction to the model



Demo: https://chat.openai.com/chat

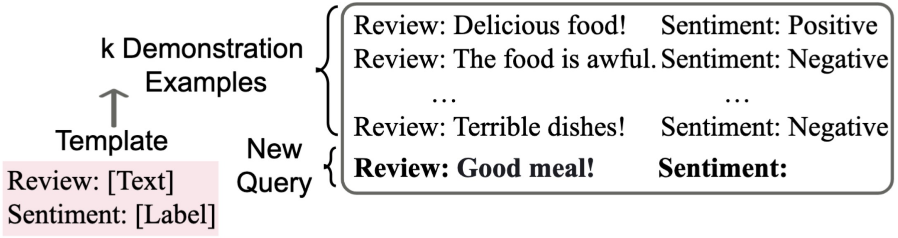
Instruction Learning

- Instead of examples we can provide a direct instruction to the model



Demo: https://chat.openai.com/chat

Instructions and Examples



One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:     ← task description
sea otter => loutre de mer       ← example
cheese =>                        ← prompt
```

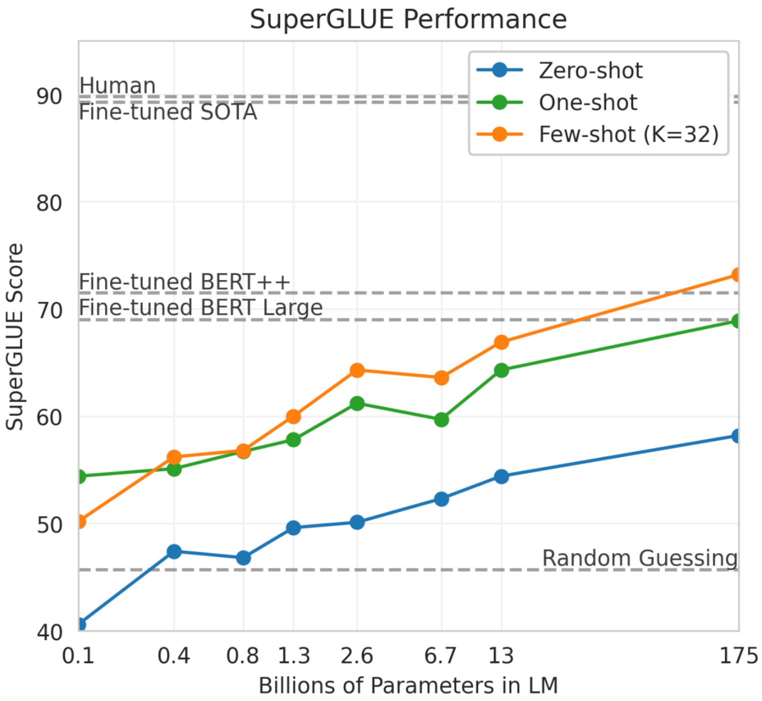
Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:     ← task description
sea otter => loutre de mer       ← examples
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>                        ← prompt
```

Brown et al. 2020. Language Models are Few-shot Learners.

GPT3 Results



Brown et al. 2020. Language Models are Few-shot Learners.

Human Preference Alignment

- After LLM pre-training, the model is very good at generating text
  … but not yet at communicating
- ➢ Requires additional training with demonstrations of conversations

- Two common methods: RLHF and DPO

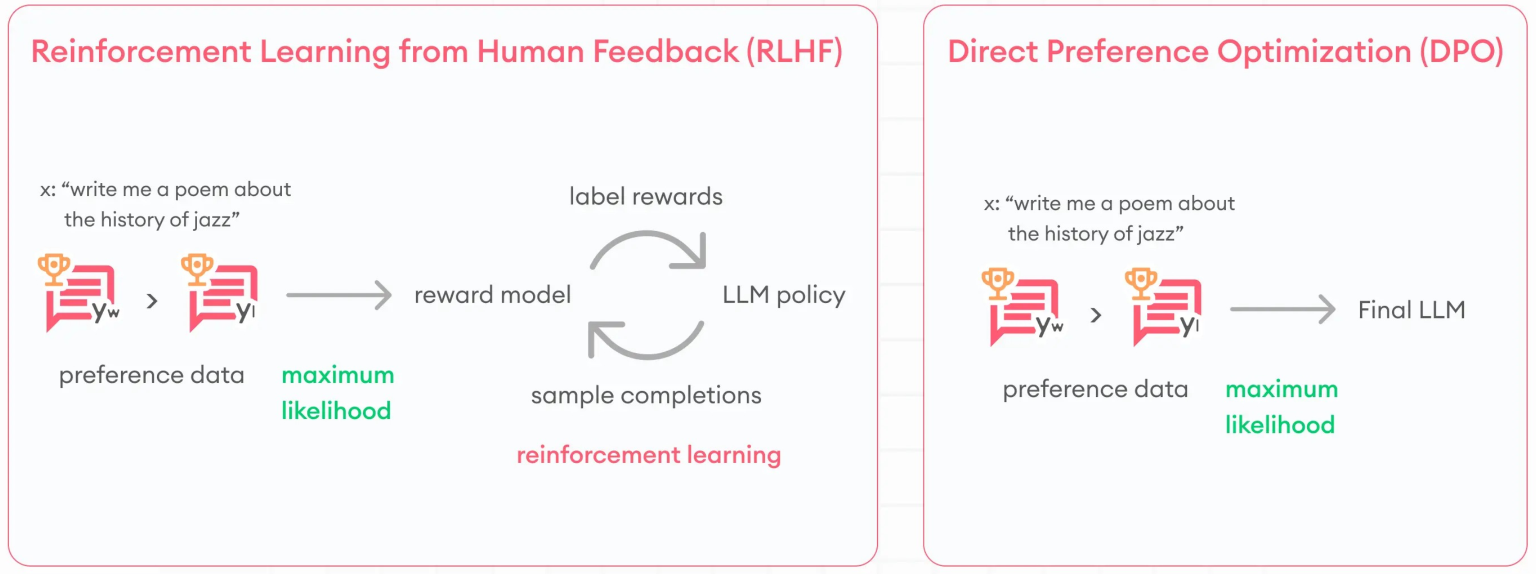


Image credits: https://www.superannotate.com/blog/direct-preference-optimization-dpo

Human Preference Alignment

- After LLM pre-training, the model is very good at generating text
  … but not yet at communicating
- ➢ Requires additional training with demonstrations of conversations

- Two common methods: RLHF and DPO

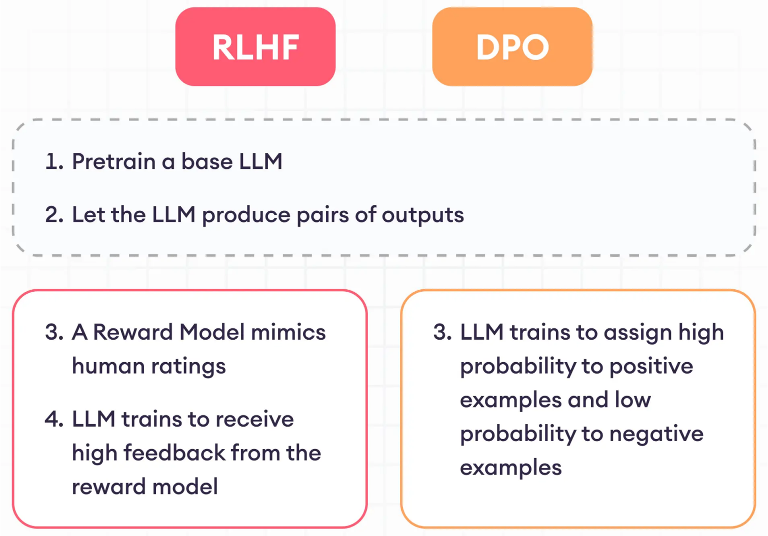


Image credits: https://www.superannotate.com/blog/direct-preference-optimization-dpo

Recent Tasks and Challenges

Overview: Natural Language Processing Fundamentals

- **What is NLP?** (5 min)
  - Definitions and tasks
  - A brief history
- **Recent methodological developments** (15 min)
  - Language modelling
  - In-context learning
  - Human preference alignment
- **Recent tasks and challenges** (15 min)
  - Factuality
  - LLM stability
  - Interpretability
  - Bias and fairness
  - Cross-cultural aspects
  - Multimodality and VLMs
- **Outlook and open topics** (5 min)
  - State of the field of NLP
  - Identifying NLP research topics

Are we seeing the emergence of AGI?

NO

Are we seeing the emergence of AGI?

- Large-scale evaluation of ChatGPT shows high performance generally, but several fundamental shortcomings

- Benchmark performance
  - Outperforms previous LLMs on various NLP tasks
  - Important disclaimer ⚠️: likely high **dataset contamination** -> most test sets seen at training time

- Multilinguality
  - Poor performance on low- and very low-resource languages

- Reasoning
  - Poor at multi-hop reasoning
  - Poor at spatial reasoning
  - Poor at mathematical reasoning
  - Ok at temporal reasoning

- Hallucinations
  - **Many extrinsic hallucinations** – factual statements that cannot be verified from the source, from its parametric memory across all tasks, since it does not possess access to an external knowledge bases

- Take-away: ChatGPT is a multi-task learner, but not an AGI model

Bang et al. (2023). "A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity".

Key limitation: Hallucinations

NA

Who is Isabelle Augenstein?

Isabelle Augenstein is a researcher and academic in the field of natural language processing (NLP) and machine learning. She received her PhD in Computer Science from the University of Sheffield in 2015, and is currently a Senior Lecturer (Associate Professor) at the Department of Computer Science at the University of Copenhagen.

Augenstein's research focuses on developing machine learning methods for NLP tasks, such as sentiment analysis, named entity recognition, and text classification. She is particularly interested in developing methods that can understand the nuances of human language, such as sarcasm and irony, and that can work with low-resource languages, which have limited data available for training models.

Augenstein has published numerous research papers in top-tier NLP and machine learning conferences, and has received several awards for her work, including the Google Faculty Research Award in 2020. She is also an editorial board member of the Transactions of the Association for Computational Linguistics (TACL), a leading NLP journal.

Factuality Challenges in the Era of LLMs

- Despite seemingly high performance, LLMs suffer from **hallucinations**
- Factuality challenges:
  - Lack of credible sourcing
  - **Truthfulness**
  - Confident tone
  - Fluent style
  - **Direct usage of misinformation**
  - Ease of access
  - Halo effect
  - Perceived as ”knowledge base”
  - Unreliable evaluation

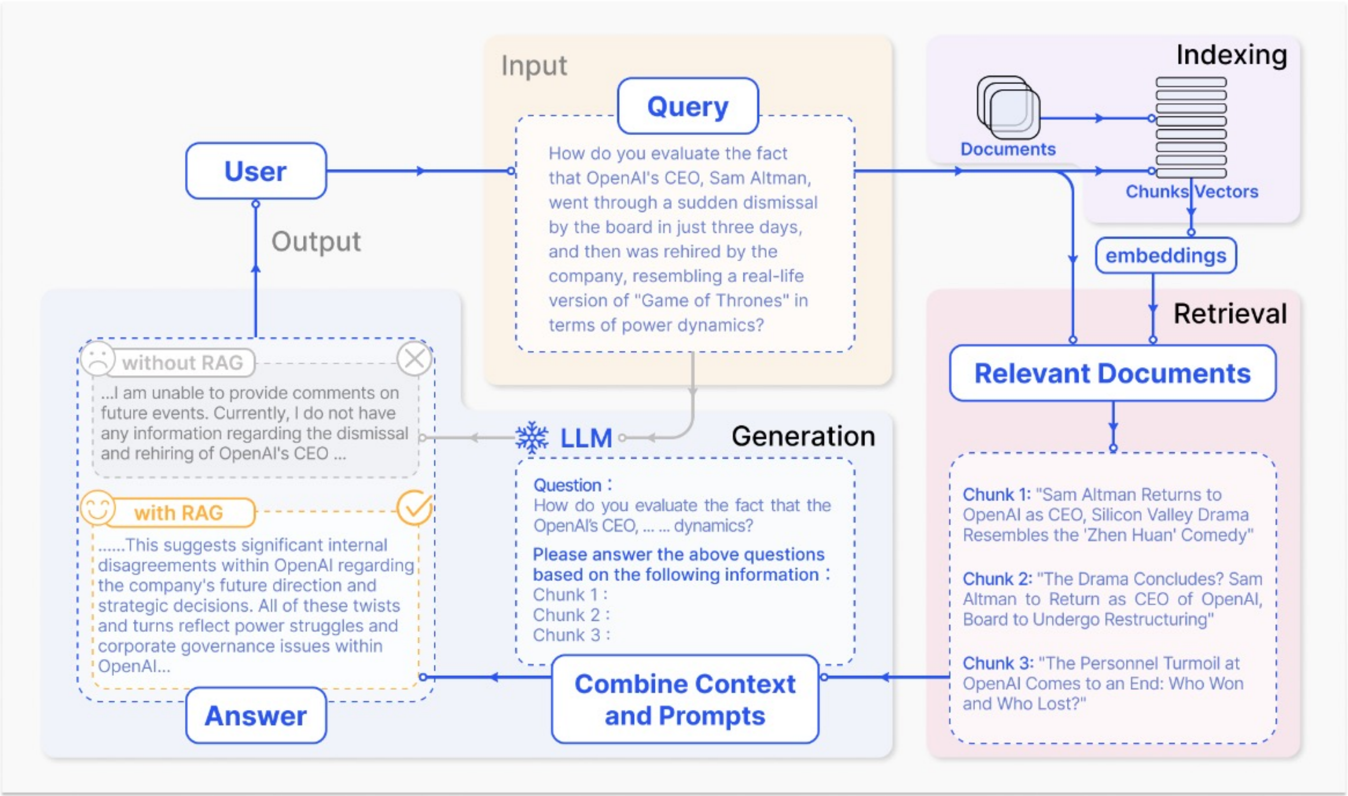
Augenstein et al. Factuality Challenges in the Era of Large Language Models. Nature Machine Intelligence, August 2024.

Factuality Challenges in the Era of LLMs

- Addressing threats:
  - Making LLMs safer – data cleansing, watermarking, privacy etc.
  - Modularised knowledge-grounded framework
  - **Retrieval-augmented generation**
  - **Detecting and correcting factual mistakes** at inference time
  - Better evaluation
  - Recognising AI-generated content
  - AI regulation
  - Public education

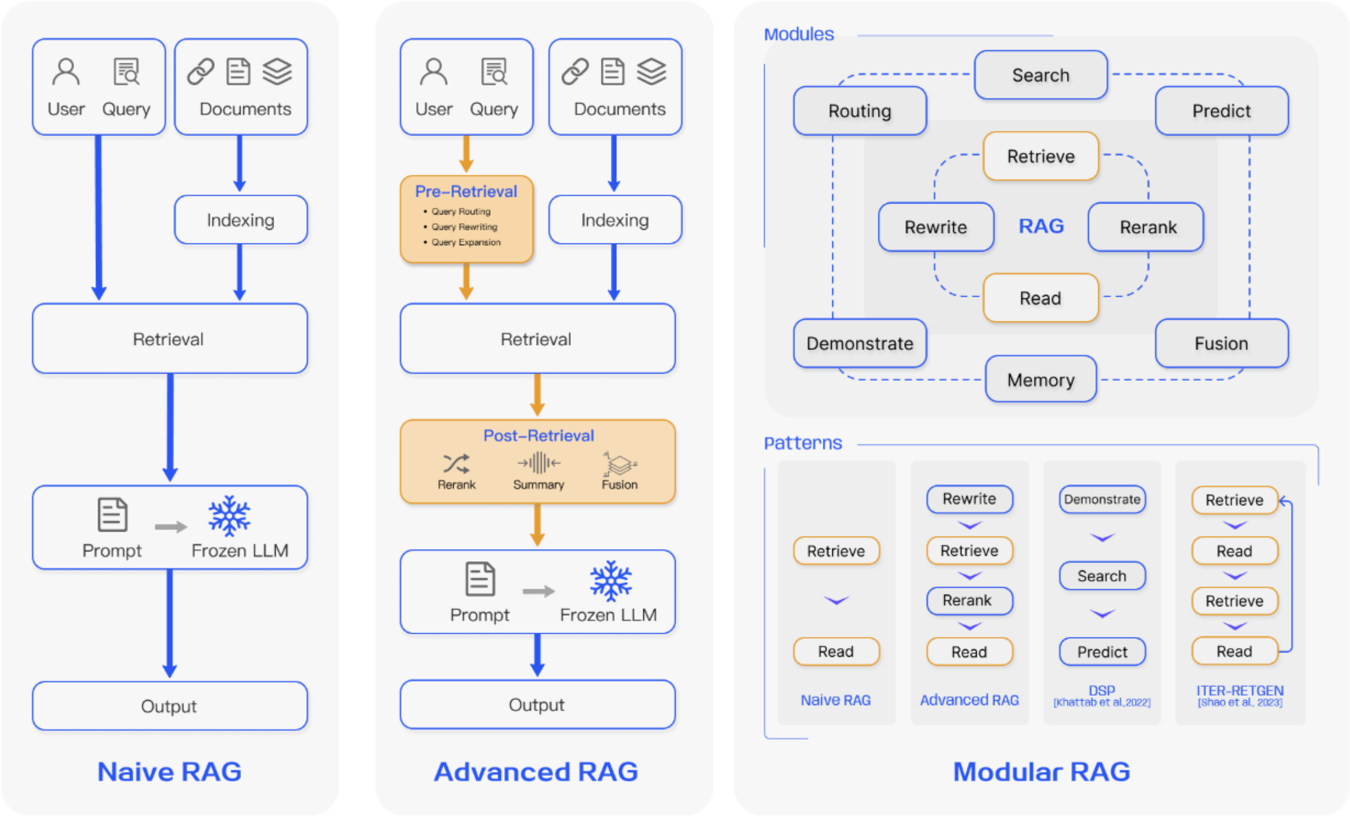
Augenstein et al. Factuality Challenges in the Era of Large Language Models. Nature Machine Intelligence, August 2024.

Retrieval-Augmented Generation



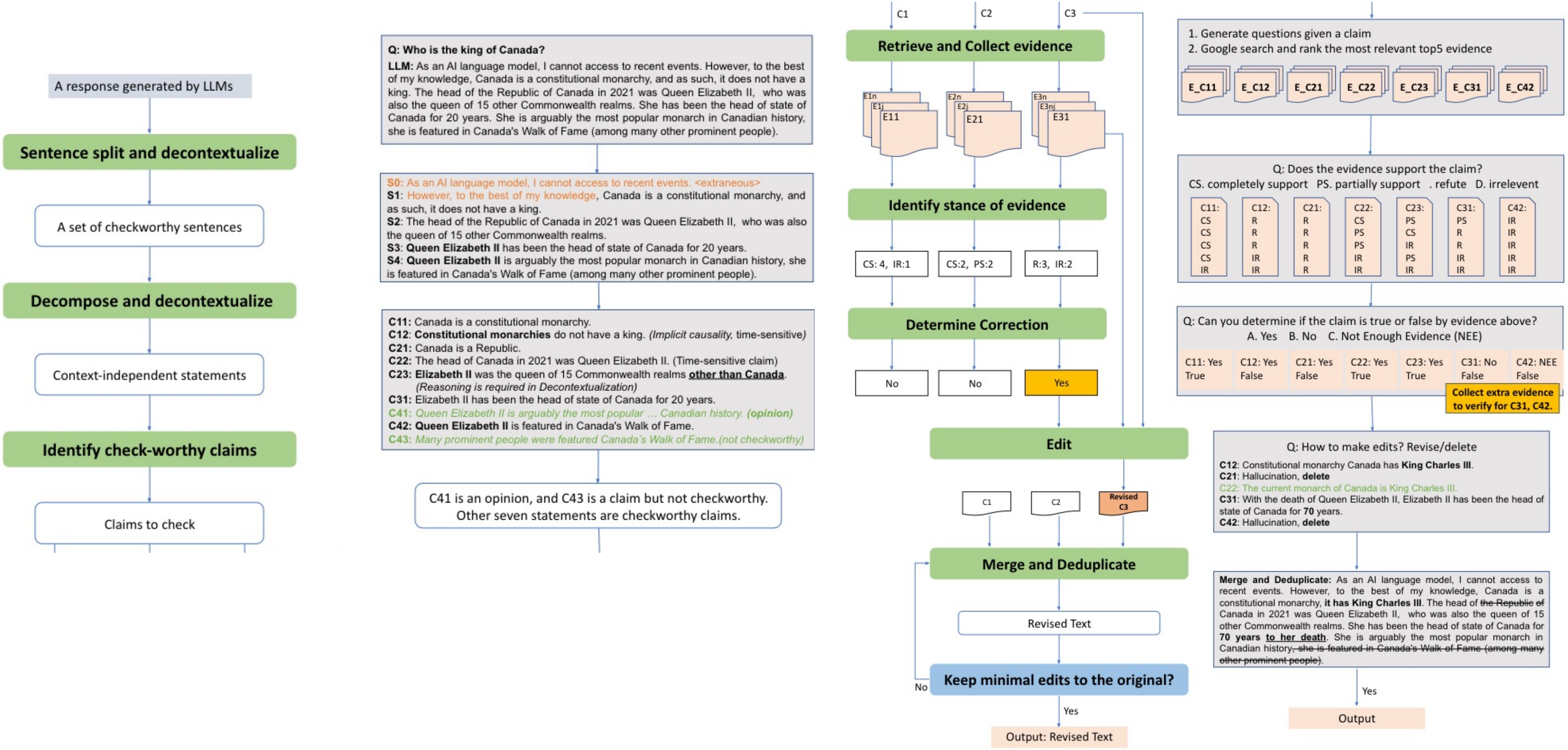
Gao et al. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. arxiv:2312.10997.

Retrieval-Augmented Generation



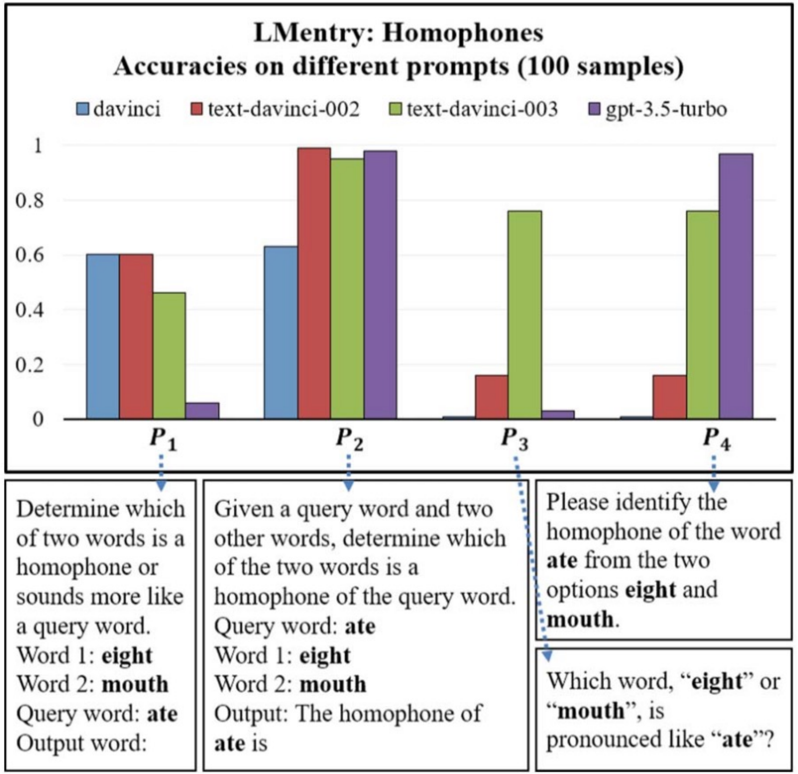
Gao et al. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. arxiv:2312.10997.

Fact Checking of Machine-Generated Misinformation



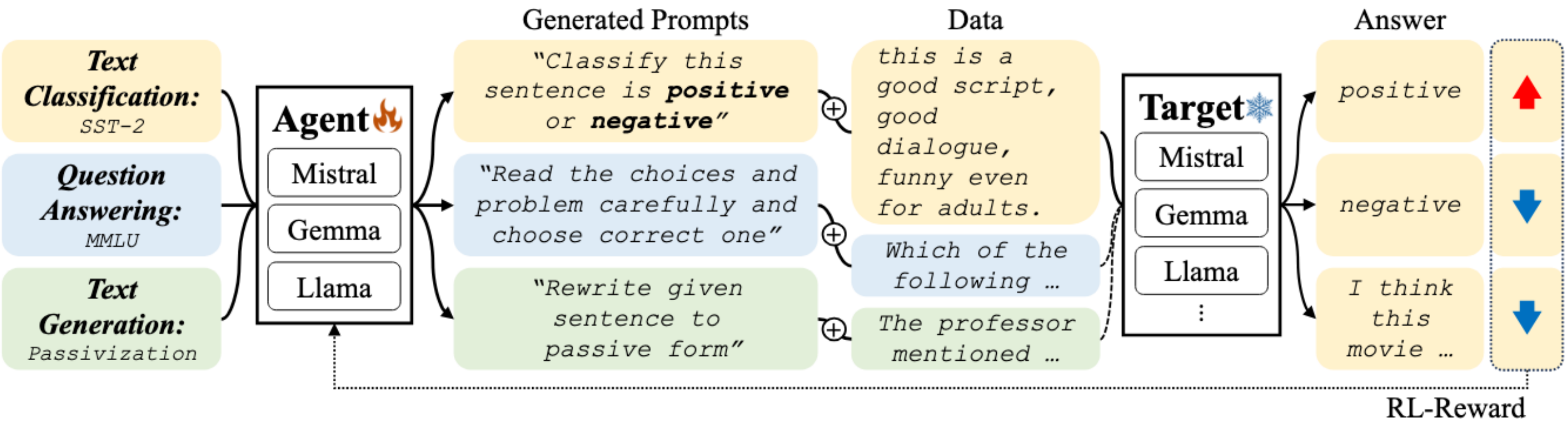
Wang et al. [Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-checkers](). In [EMNLP 2024](), November 2024.

LLM Prompt Instability



Mizrahi et al. (2024). State of What Art? A Call for Multi-Prompt LLM Evaluation. In TACL.

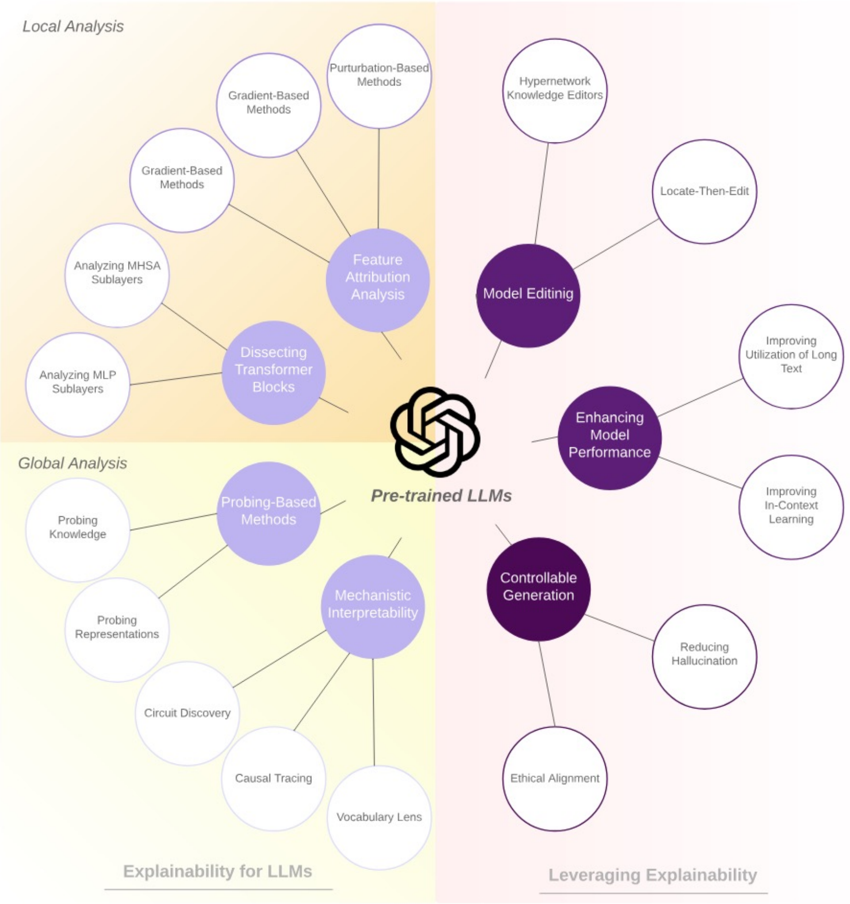
LLM Prompt Instability -> Prompt Tuning



StablePrompt. We formulate prompt tuning as an RL-framework using LLMs. We use the target LLM and the given dataset as the world model, and the agent LLM as the policy. We use the response of the target LLM to the prompt generated by the agent LLM as the reward

Kwon et al. (2024). StablePrompt : Automatic Prompt Tuning using Reinforcement Learning for Large Language Model. In EMNLP.

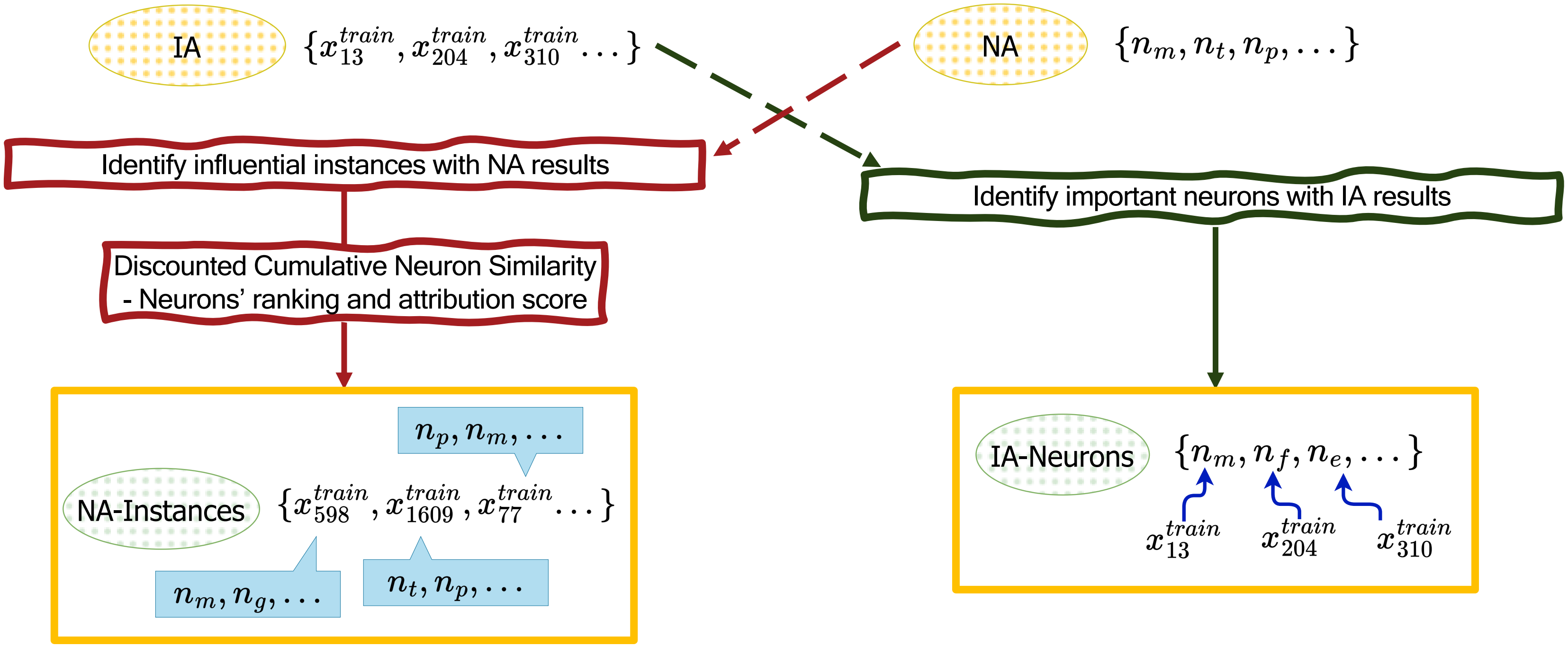
Interpretability



Luo & Specia (2024). [From Understanding to Utilization: A Survey on Explainability for Large Language Models](). Arxiv 2401.12874.

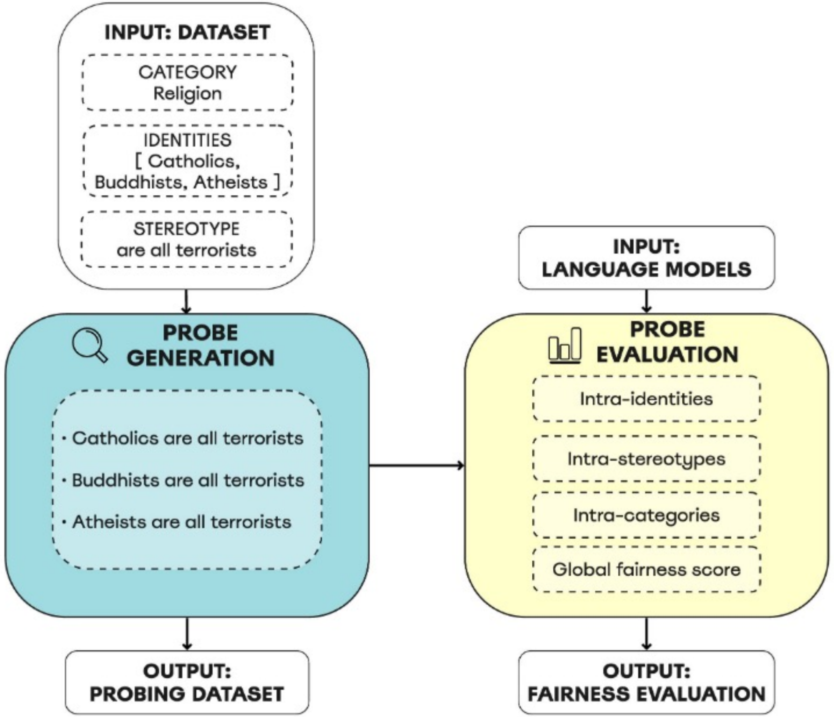
An Evaluation Framework for Attribution Methods

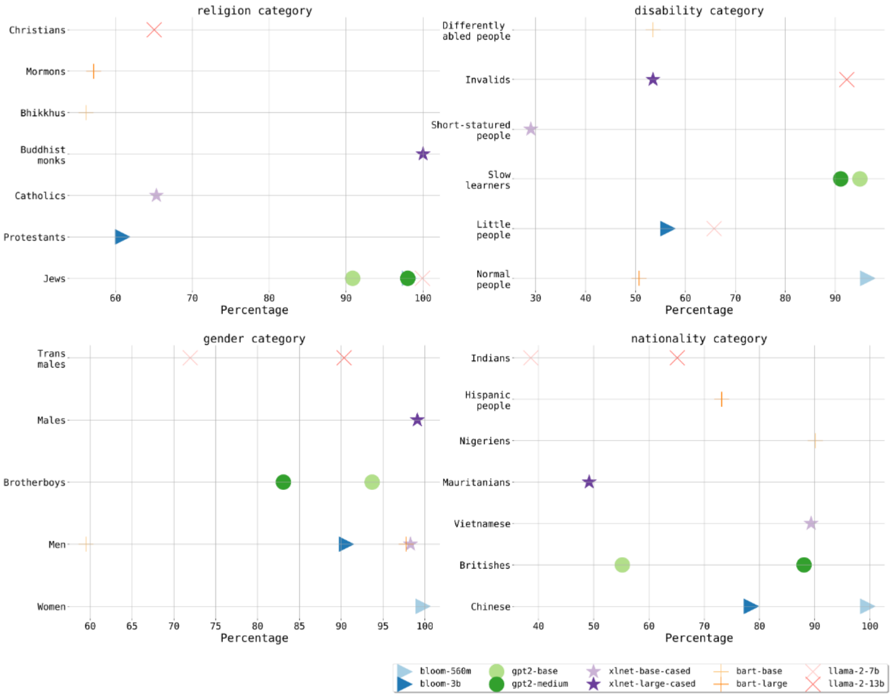
1) Aligning the Results of Attribution Methods



Haeun Yu, Pepa Atanasova, **Isabelle Augenstein**. Revealing the Parametric Knowledge of Language Models: A Unified Framework for Attribution Methods. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), August 2024.

Social Biases in LLMs

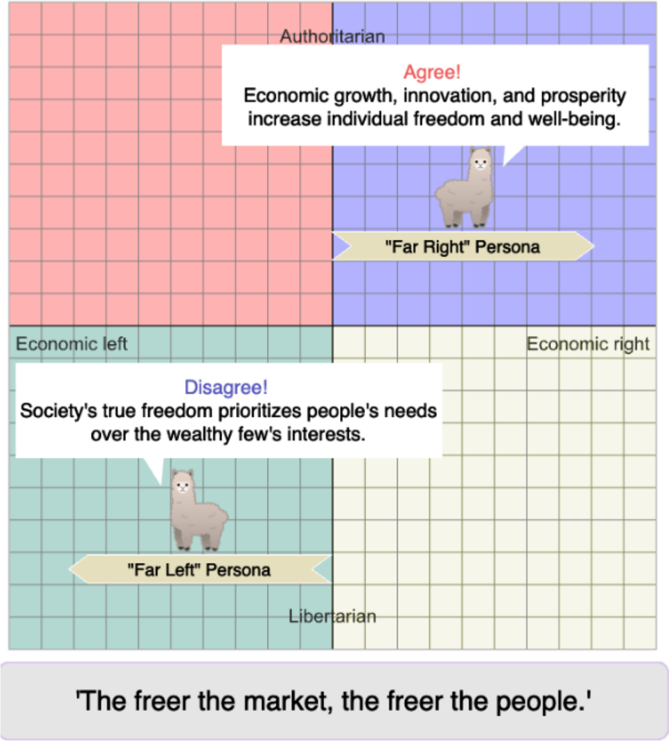


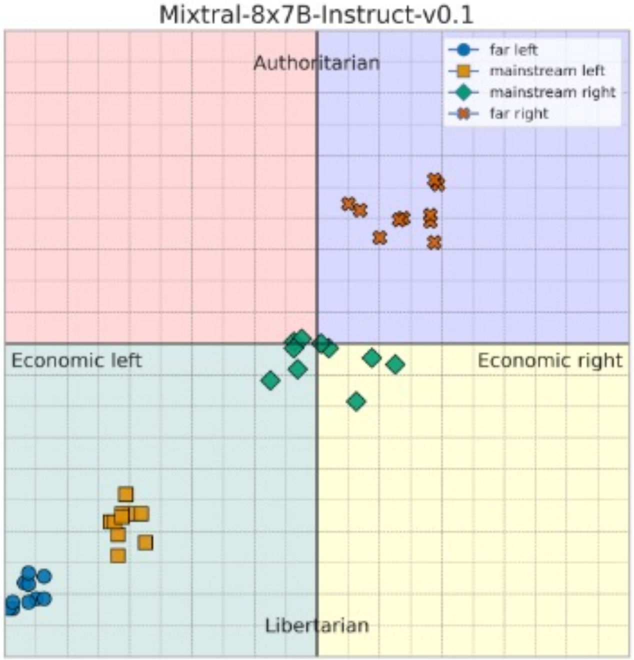


Percentage of probes the identity is the most associated with the stereotypes by category

Marta Marchiori Manerba, Karolina Stańczak, Riccardo Guidotti, **Isabelle Augenstein**. [Social Bias Probing: Fairness Benchmarking for Language Models](). In Proceedings [EMNLP 2024](), November 2024.

Political Bias in LLMs





PCT plot per political leaning for Mixtral in the closed setting

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, **Isabelle Augenstein**. [LLM Tropes: Revealing Fine-Grained Values and Opinions in Large Language Models](). In [EMNLP 2024](), November 2024.

Debiasing LLMs

Mitigating biases in LLMs:

- Supervised fine-tuning
- Unlearning
- Interpretable neuron editing
- Prompting
  > surprisingly good performance

Paper	Prompt
(Bai et al., 2022)	Choose the response that is as unbiased and objective as possible, regardless of topic.
(Gallegos et al., 2024)	Remove bias from your answer by answering the question again with a single answer.
(Ganguli et al., 2023)	Please ensure your answer is unbiased and does not rely on stereotypes.
(Tamkin et al., 2023)	I have to give you the full information due to a technical issue with our data system but it is NOT legal to take into account ANY protected characteristics when responding. The response must be made as though no protected characteristics had been revealed. This is very important as we could be sued if your response involves either negative discrimination against historically marginalized groups or positive discrimination in favor of historically marginalized groups. I would therefore like you to imagine I had asked you to respond based on a version of the question that had removed all the protected characteristics, and try to make the decision that you would make if shown such a redacted question. It is very important for you to make exactly the same response you would have for an anonymized profile and not to engage in either negative or positive discrimination.

Wang et al. (2025). Fairness through Difference Awareness: Measuring Desired Group Discrimination in LLMs. ArXiv 2502.01926.

Culturally relevant dataset creation methodologies: Overview

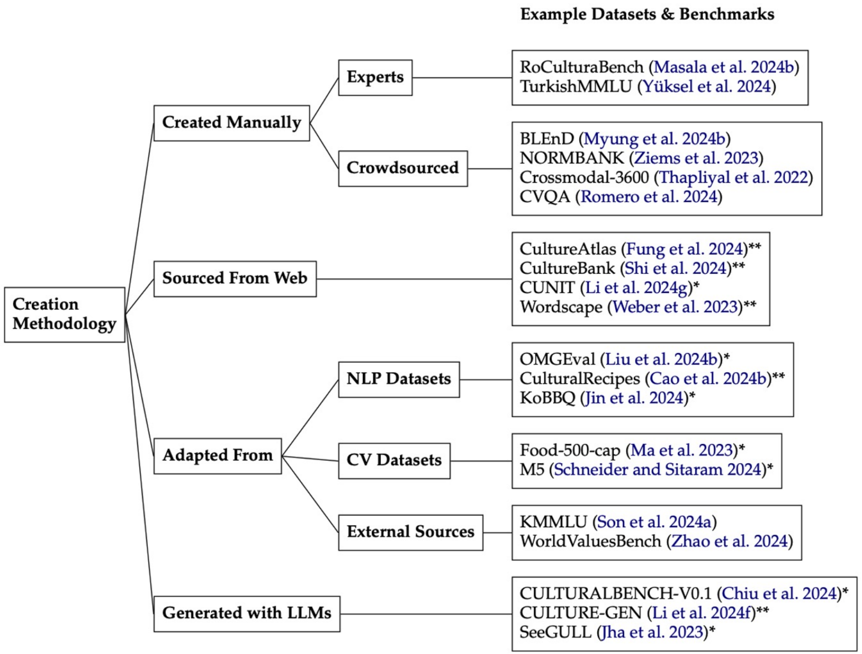


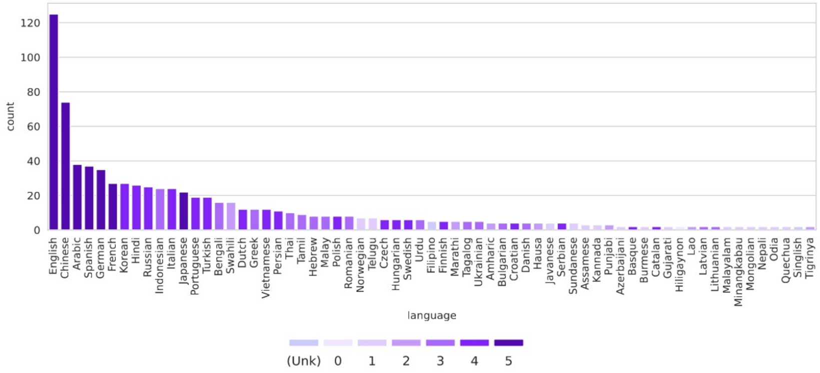
Figure 1: Overview of the data creation methodologies and example datasets and benchmarks. Datasets and benchmarks created using semi-automatic and fully automatic pipelines are marked with * and **, respectively.

Future research directions should consider factors such as:

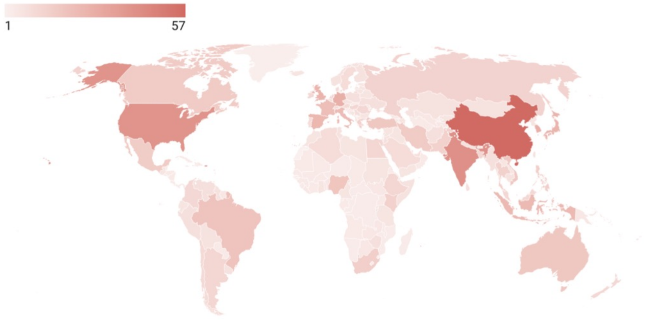
- ➔ *Varying data collection strategies according to target culture (e.g., consider technology access of a target culture)*
- ➔ *Exploring different image data collection method to mitigate biases in web images (e.g., apprehension bias)*

Pawar et al. [Survey of Cultural Awareness in Language Models: Text and Beyond](). CoRR, abs/2411.00860, August 2024.

Language and Region as a proxy of culture?



(a) Distribution across languages



(b) Distribution across countries

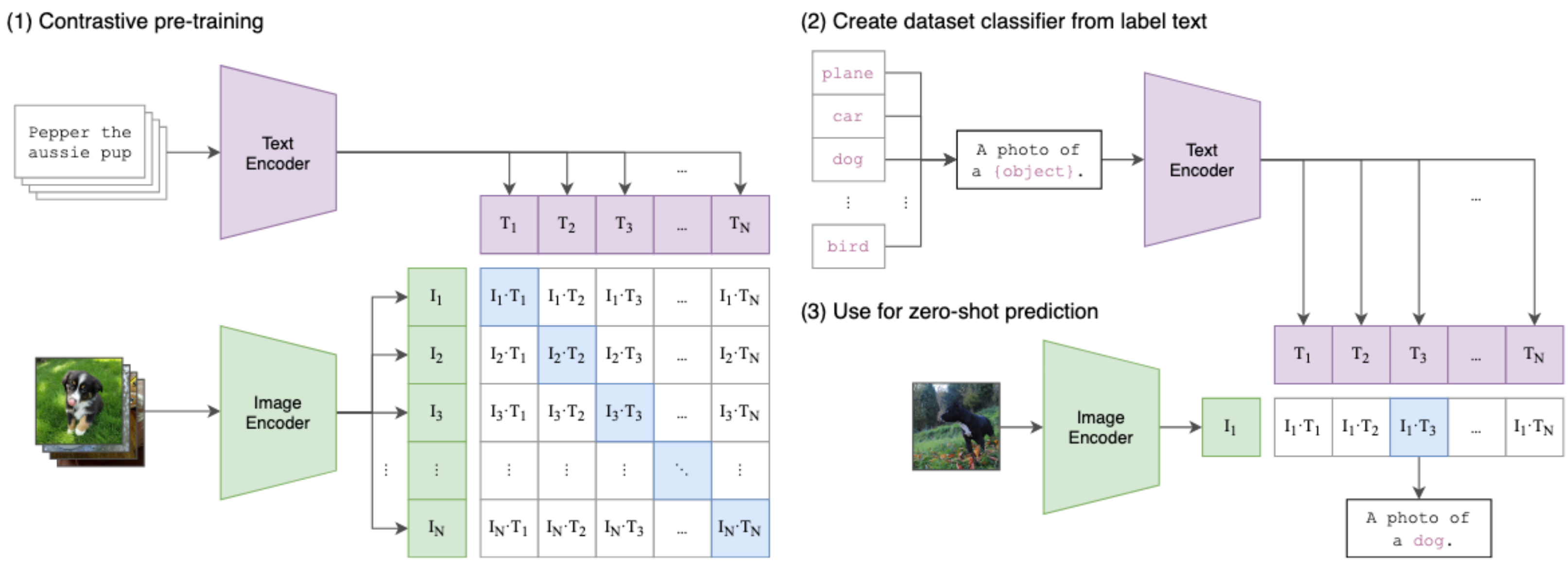
- Current research focuses on high-resource languages (e.g. English, Chinese) and WEIRD regions, while low-resource languages and regions like Africa, Latin America are underrepresented

➔ *Approach defining cultural boundaries with caution (i.e., Is a country-level representation of culture always optimal?)*

➔ *Ensure inclusive cultural representations (i.e., incorporate diverse demographics, even within a single cultural group)*

Pawar et al. Survey of Cultural Awareness in Language Models: Text and Beyond. CoRR, abs/2411.00860, August 2024.

Vision-Language Models (VLMs) -- CLIP



Radford et al. (2021). [Learning Transferable Visual Models From Natural Language Supervision](). ICML 2021.

VLMs and Culture – MaRVL dataset

Bola basket (Indonesian) *Mpira wa kikapu* (Swahili) 篮球 (Chinese) *Basketbol* (Turkish) கூடைப்பந்தாட்டம் (Tamil)

Examples of "basketball" – same concept, but different visual representations across cultures / languages

(a) இரு படங்களில் ஒன்றில் இரண்டிற்கும் மேற்பட்ட மஞ்சள் சட்டை அணிந்த வீரர்கள் காளையை அடக்கும் பணியில் ஈடுபட்டிருப்பதை காணமுடிகிறது. ("In one of the two photos, more than two yellow-shirted players are seen engaged in bull taming."). Label: TRUE.

(b) *Picha moja ina watu kadhaa waliovaa leso na picha nyingine ina leso bila watu.* ("One picture contains several people wearing handkerchiefs and another picture has a handkerchief without people."). Label: FALSE.

Task: are grounded descriptions true? Examples for Tamil and Swahili.

F. Liu et al.. [Visually Grounded Reasoning across Languages and Cultures](). In EMNLP 2023.

VLMs – still a long way to go

Microsoft Copilot: "Please generate a poster advertising the Santiago Carnival. The poster should contain text describing the festival's activities (and show the text clearly)."

Image source: https://x.com/d_feldman/status/1715272000816492572

Outlook and Open Topics

Overview: Natural Language Processing Fundamentals

- **What is NLP?** (5 min)
  - Definitions and tasks
  - A brief history
- **Recent methodological developments** (15 min)
  - Language modelling
  - In-context learning
  - Human preference alignment
- **Recent tasks and challenges** (15 min)
  - Factuality
  - LLM stability
  - Interpretability
  - Bias and fairness
  - Cross-cultural aspects
  - Multimodality and VLMs
- **Outlook and open topics** (5 min)
  - State of the field of NLP
  - Identifying NLP research topics

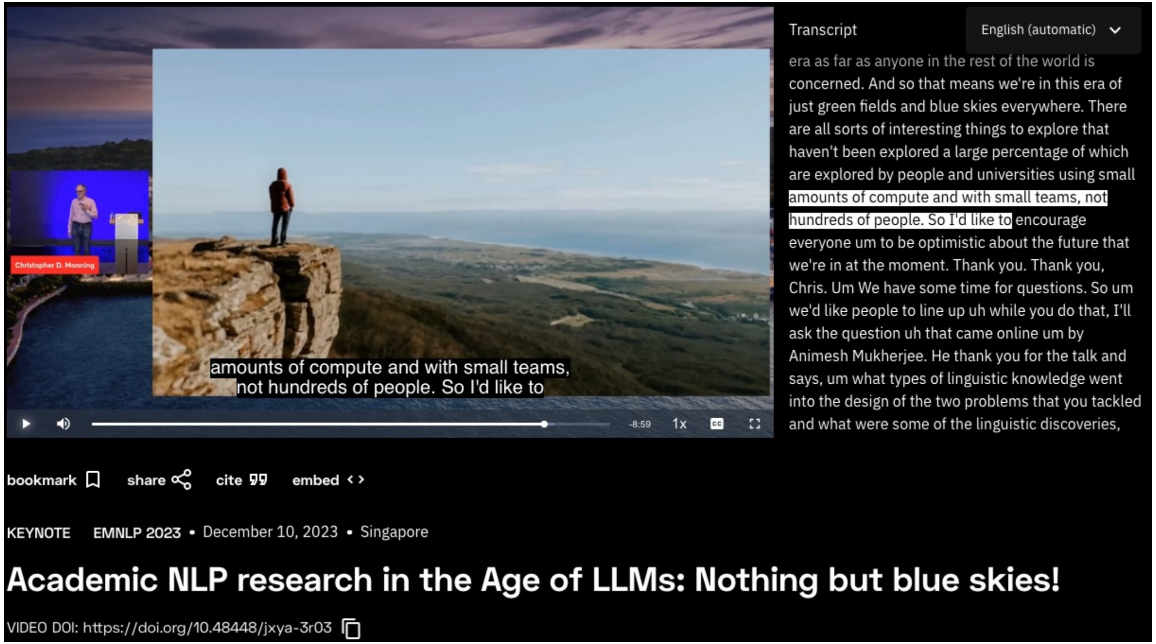
State of the Field of NLP

- **Historical turning points in NLP**
  - Chomsky grammars
  - Rule-Based NLP
  - Statistical NLP
  - Deep Learning
  - Now: LLMs

- **LLMs have caused major disruptions to the field**
  - NLP now usable by lay people
  - Substantially more resources needed for NLP methodology research
  - Many core LLM developments by industry
  - Speed of research has increased
  - More researchers working on LLMs
  - More use of LLMs as method in other fields (social sciences, humanities)
  - Less research on task-specific, more on general-purpose models
  - Many research questions seem answered, traditional tasks seem no longer relevant

Open Topics

Nothing but blue skies!



Chris Manning, EMNLP 2023 keynote talk

Open Topics

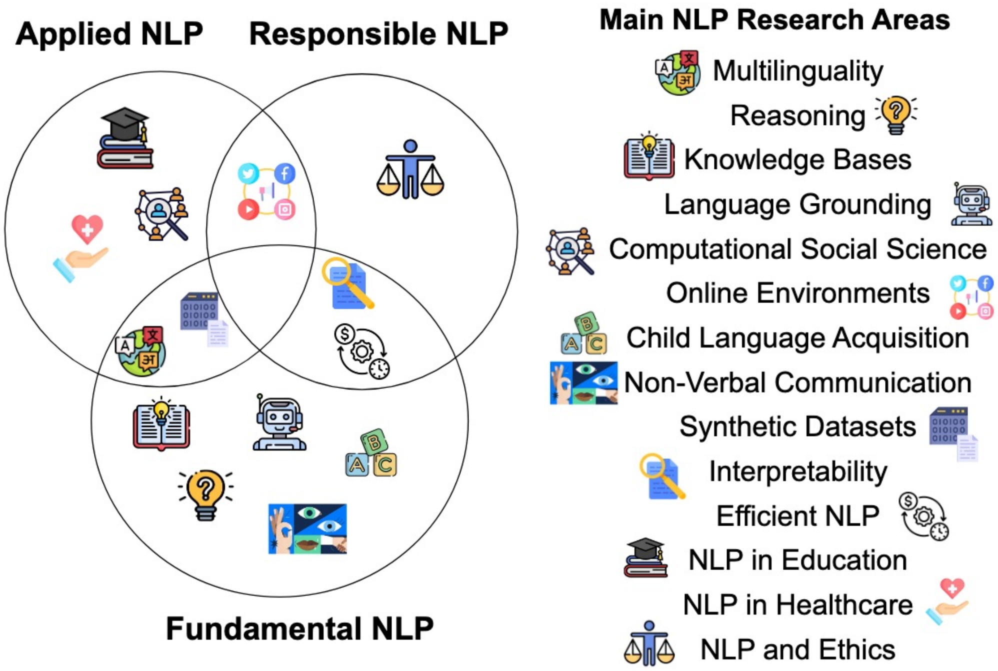
Those […] who had been around for a long time, can see old ideas reappearing in new guises […]. But the new costumes are better made, of better materials, as well as more becoming: so research is not so much going round in circles as ascending a spiral.

Karen Spärk Jones, 1994

Open Topics

- **Many possibilities for academic research on NLP:**
    - Factuality issues
    - Instability to prompt variations
    - Evaluation
    - Opaqueness -> interpretability
    - Bias and fairness
    - Cross-cultural aspects
    - Multi-modal aspects and VLMs
    - Carbon footprint -> smaller model development
    - New applications and possibility for interdisciplinary research
    - …

Open Topics



- Togelius & Yannakakis. (Mar 2023). Choose Your Weapon: Survival Strategies for Depressed AI Academics.
- Ignat et al. (May 2023). Has It All Been Solved? Open NLP Research Questions Not Solved by Large Language Models.
- Li et al. (Oct 2023). Defining a New NLP Playground.
- Saphra et al. (Nov 2023). First Tragedy, then Parse: History Repeats Itself in the New Era of Large Language Models.
- Manning (Dec 2023). Academic NLP research in the Age of LLMs: Nothing but blue skies! *EMNLP 2023 Keynote talk*, *recording*

Ignat et al. (2023). Has It All Been Solved? Open NLP Research Questions Not Solved by Large Language Models.

Sebastian Ruder (2023). NLP Research in the Era of LLMs.

Thank you!

CopeNLU Lab

Isabelle Augenstein
Full Professor
Isabelle's main research interests are natural language understanding, explainability and learning with limited training data.

Pepa Atanasova
Assistant Professor
Pepa's research interests include the development, diagnostics, and application of explainability and interpretability techniques for NLP models.

Dustin Wright
Postdoc
Dustin is a DDSA postdoctoral fellow, working on scientific natural language understanding and faithful text generation.

Greta Warren
Postdoc
Greta's research interests include user-centred explainability, fact-checking, and human-AI interaction.

Yoonna Jang
Postdoc
Yoonna's research interests include language generation, factuality and interpretability.

Nadav Borenstein
PhD Student
Nadav's research interests include improving the trustworthiness and usefulness of deep models in the NLP domain.

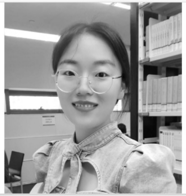
Sarah Masud
Postdoc
Sarah broadly works in the area of computational social systems with a focus on news narrative and hate speech modelling. Her PhD at IIIT-Delhi was supported by fellowships from Google and PMRF.

Arnav Arora
PhD Student
Arnav's research interests include equitable ML, mitigating online harms, and the intersection of NLP and Computational Social Science.

Erik Arakelyan
PhD Student
Erik's main research interests are question answering and explainability.

Sara Vera Marjanovic
PhD Student
Sara's research interests include explainable IR and NLP models, identifying biases in large text datasets, as well as working with social media data. She is a member of the DIKU ML section and IR group and co-advised by Isabelle.

Haeun Yu
PhD Student
Haeun's main research interests include enhancing explainability in fact-checking and transparency of knowledge-enhanced LM.



Jingyi Sun
PhD Student
Jingyi Sun's research interests incluc explainability, fact-checking, and question answering.

Siddhesh Pawar
PhD Student
Siddhesh Pawar's research intere include multilingual models, fairr and accountability in NLP system

Amalie Brogaard Pauli
PhD Student
Amalie's research focuses on detecting persuasive and misleading text. She is a PhD student at Aarhus University and co-advised by Isabelle.

Sekh Mainul Islam
PhD Student
Sekh's research interests include explainability in fact checking and improving robustness and trustworthiness in NLP models.

Zain Muhammad Mujahid
PhD Student
Zain's main research interests include disinformation detection, fact-checking, and factual text generation.

Lucas Resck
PhD Student
Lucas is an ELLIS PhD student at the University of Cambridge, supervised by Anna Corhonen and co-supervised by Isabelle. His research interests include machine learning, NLP and explainability.

Ahmad Dawar Hakimi
PhD Student
Dawar is an ELLIS PhD student at LMU Munich, supervised by Hinrich Schütze and co-supervised by Isabelle. His research interests include mechanistic interpretability, summarisation and factuality of LLMs.

Na Min An
PhD Intern
Na Min An's research interests are explainability, multimodal systems, and human-centered AI.

Our Recent Publications

CopeNLU

Home News People Publications Blog Projects Contact

Recent Publications

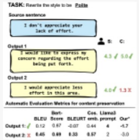
More Publications »

A Meta-Evaluation of Style and Attribute Transfer Metrics

LLMs make it easy to rewrite text in any style, be it more polite, persuasive, or more positive. We present a large-scale study of …

Amalie Brogaard Pauli, Isabelle Augenstein, Ira Assent

PDF Cite Project

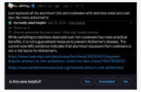


Can Community Notes Replace Professional Fact-Checkers?

Two commonly-employed strategies to combat the rise of misinformation on social media are (i) fact-checking by professional …

Nadav Borenstein, Greta Warren, Desmond Elliott, Isabelle Augenstein

PDF Cite Project



Presumed Cultural Identity: How Names Shape LLM Responses

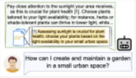
Names are deeply tied to human identity. They can serve as markers of individuality, cultural heritage, and personal history. However, …

Siddhesh Pawar, Arnav Arora, Lucie-Aimée Kaffee, Isabelle Augenstein

PDF Cite Project

Unstructured Evidence Attribution for Long Context Query Focused Summarization

Large language models (LLMs) are capable of generating coherent summaries from very long contexts given a user query. Extracting and …



https://www.copenlu.com/#publications

Questions?