



Natural Language Processing

COM6513/4513 Natural Language Processing

Nafise Sadat Moosavi

n.s.moosavi@sheffield.ac.uk

Computer Science School

Lecture Plan

1. The module
2. Introduction to NLP
 - a. Inherent challenges of understanding human language
 - b. Past and present recap
 - c. Open challenges in NLP
3. Representing Text as Vectors
 - a. Recap
 - b. Evaluation

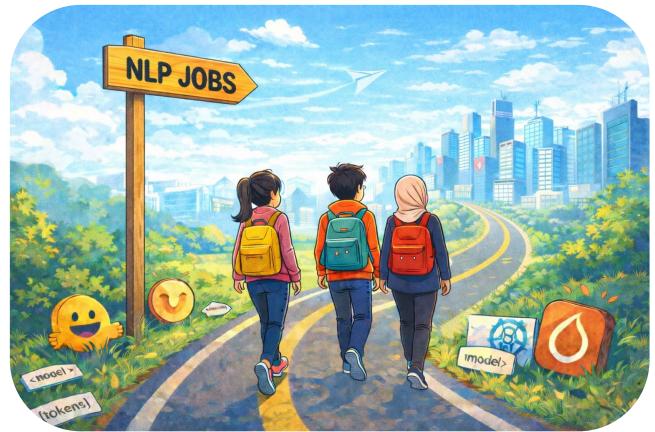
About

- Lectures: 11-13 Tuesdays
- Labs: 11 am - 12:00 pm Thursdays
 - Atsuki Yamaguchi
 - Feng He
 - Samuel Lewis-Lim
- Diamond, Computer Room 06



About

- Lectures:
 - Why NLP suddenly works
 - What's inside modern language models
- Labs:
 - Hands-on with real NLP models
 - Running and adapting modern language models



About

Lab sessions:

- Hands-on coding and experimentation, aligned with lectures
- Sessions are 1h long
- It's not expected to come up with a full solution in 1h. You can work on the lab assignment at home.
- Lab sessions are not marked and solutions will be provided.

About

- Lectures, Lab content, announcements, general questions
 - BlackBoard
- Assessment
 - 70% exam
 - where everything is assessed: lecture slides, bibliographical references, classroom discussion, etc.
 - 30% Assignment:
 - Hand-out: 17 Mar 2026
 - Hand-in: 15:00 on 2 Apr 2026
 - Review carefully rules regarding the use of unfair means!

Feedback

To you:

- During the lab sessions
- Blackboard Discussions
- Assignment

And to us:

- NSS evaluation
- Module evaluation
- Email

If something is unclear, or particularly helpful, we want to know.

Not just when things go very wrong (or very right!).

Feedback

- Accessibility
- Difficulty level
- Things you would like to learn
- Things you didn't like to learn :)
- ...

This year's module has been substantially updated, informed by student feedback and recent developments in NLP.

What do we hope to teach?

- Foundations of modern language models
 - From classical representations and n-grams to neural and Transformer-based LMs
 - Tokenization, training objectives, and evaluation
- How contemporary NLP systems work
 - Encoder, decoder, and encoder–decoder architectures
 - Pretraining, fine-tuning, chain-of-thought
- Practical skills for working with NLP models
 - Using and adapting pretrained models in PyTorch and Hugging Face
 - Fine-tuning, parameter-efficient methods, and prompting
 - Analysing, interpreting, and evaluating model behaviour

Course Bibliography

Jurafsky and Martin. Speech and Language Processing, Prentice Hall [[3rd edition](#)]

Jacob Eisenstein. 2019. Introduction to Natural Language Processing. MIT Press.
(A draft can be found [here](#))

other materials referenced at the end of each lecture

Attendance Monitoring

VR-TB-CA



Natural Language Processing



1

Go to wooclap.com

2

Enter the event code in the top banner



Enable answers by SMS

Event code
URZSAR

Natural Language Processing

- Enabling computers to understand, interpret, and produce human languages
 - Bridging the gap between human communication and computer understanding



Key Goals

- **Understanding**

Extract meaning, sentiment, and intent from language

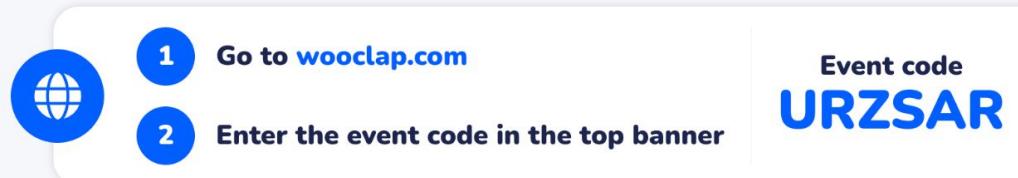
- **Generation**

Produce fluent and contextually appropriate text

- **Interaction**

Enable natural communication between humans and machines

What are the Challenges?



Enable answers by SMS

Challenges

- Complexity of human language
 - Slang
 - Idioms
 - Dialects
- Ambiguity and Context
 - Understanding context and disambiguating meanings

Challenges

- Complexity of human language
 - British English: We need to table the discussion on budget allocation until next meeting
 - American English: Let's table the discussion about the new project for now

Word Ambiguity

After withdrawing from the bank, she sat by the river bank

Word Ambiguity

After withdrawing from the bank, she sat by the river bank



Pronoun Resolution

John told Peter that he would have to leave the office by 5 PM

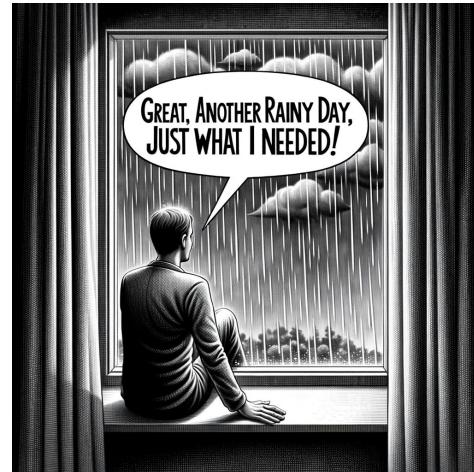


Idiomatic Expression

When he finally kicked the bucket, his family didn't know how to handle the farm

Sarcasm and Irony

Great, another rainy day. Just what I needed!



Polysemy: Words with Multiple Meanings

He was interested in the match

Polysemy

He was interested in the match



Figurative Language: Metaphors and Similes

Time is a thief



Ellipsis and Implied Information

Want to go for coffee?

Love to, but swamped

Language Evolution and Neologisms

- Language is constantly evolving
 - COVID-19
 - Zoom
 - Ghosting
 - Cryptocurrency
 - Phubbing
 - Meme
 - Crowdfunding

Cultural References and Allusions

He's the Sherlock Holmes of programmers

Cultural References and Allusions

They're like The Butterfly Lovers

In Chinese culture:

- A happy love story?
- A tragic loss?

Ambiguous and Complex Questions

Can you tell me the way to San Jose?

- 1 meaning?
- 2?
- more?

NLP Applications



Event code
QLLRKP

1 Go to wooclap.com
2 Enter the event code in
the top banner

SMS Enable answers by SMS

A call-to-action card from Wooclap. It features a globe icon, step-by-step instructions for joining the event online, and the event's unique identifier, "QLLRKP". A "Enable answers by SMS" button is also present at the bottom.

Applications

Machine Translation

Welcome to our first lecture

Bienvenue à notre première
conférence

به اولین کلاس خوش امدید

Willkommen zu unserem ersten
Vortrag

欢迎大家来到我们的第一堂课

أهلاً بكم في محاضرتنا الأولى

हमारे पहले व्याख्यान में आपका
स्वागत है

Chatbots and Virtual Assistants



Automated Customer Support

Chatbots to enhance customer experience

- Instant responses to queries
- Reducing wait times
- Freeing human agents to handle more complex issues



Sentiment Analysis for Market Research

- Analyzing sentiments in social media, reviews, and feedback
- Measuring public opinion, brand reputation, and customer satisfaction
- Enabling data-driven decision-making



Content Recommendation Systems

- Analyzing user preferences, search queries, views or purchase history
- Improving user experience and engagement
- Netflix, YouTube, and Amazon



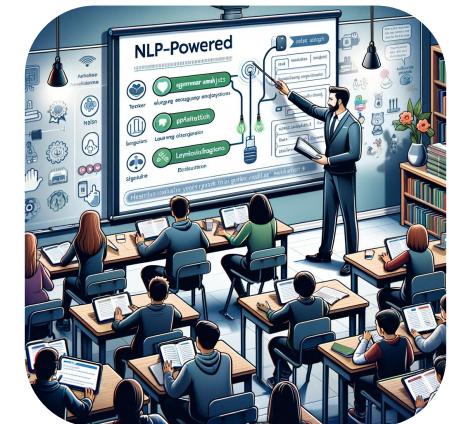
Automatic Text Summarization

- Concise summaries of long documents, reports, or articles
- Quickly assimilating information from multiple sources



Educational Tools and Learning Aids

- Real-time feedback, grammar assistance, and personalized learning experiences
- Enhancing education accessibility and effectiveness



NLP History

Rule-based systems

0 Paul Allen was born on January 21 , 1953 , in Seattle , Washington . 0 Allen attended
3 Lakeside School , where 0 he befriended 1 Bill Gates . 2 0 Paul and 1 Bill used a
teletype terminal at 3 2 their high school , Lakeside , to develop 2 their programming skills
on several time - sharing computer systems .

Pass	Type	Features
1	N	exact extent match
2	N,P	appositive predicate nominative role appositive relative pronoun acronym demonym
3	N	cluster head match & word inclusion & compatible modifiers only & not i-within-i
4	N	cluster head match & word inclusion & not i-within-i
5	N	cluster head match & compatible modifiers only & not i-within-i
6	N	relaxed cluster head match & word inclusion & not i-within-i
7	P	pronoun match

NLP History

Statistical NLP



Feature
extraction

Machine
Learning



NLP History

Statistical NLP



Feature
extraction

Machine
Learning



Chatbots & Virtual Assistant

NLP History

Statistical NLP



Feature
extraction

Machine
Learning



NLP History

Statistical NLP



Feature
extraction

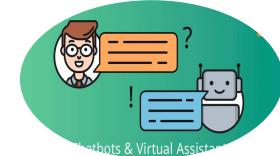
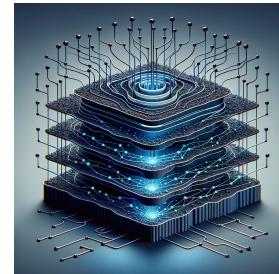
Machine
Learning



NLP History

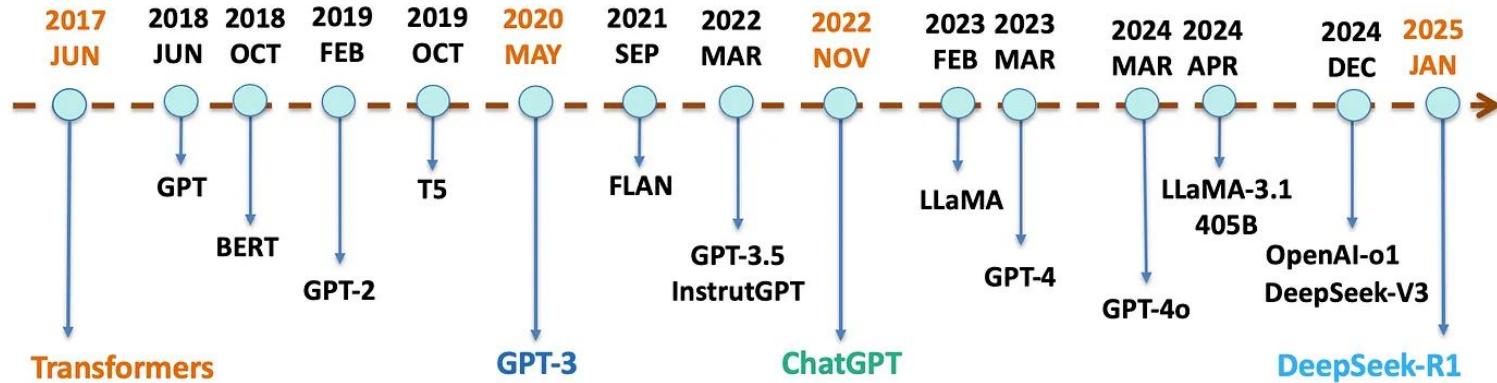
Neural networks and deep learning

- Word embedding
 - Enhanced input representation
- Transformer architectures and pre-training
 - Enhanced language understanding



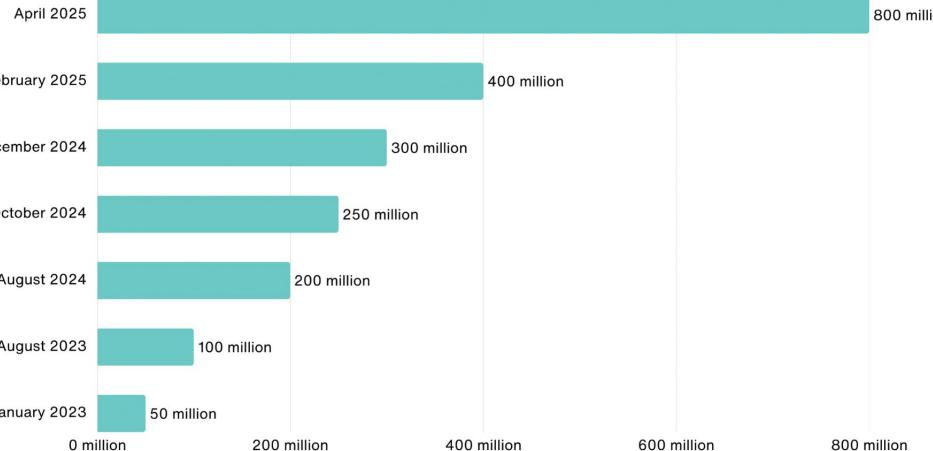
OpenAI's ChatGPT

A Brief History of LLMs



Source: <https://medium.com/@lmpo/a-brief-history-of-lmms-from-transformers-2017-to-deepseek-r1-2025-dae75dd3f59a>

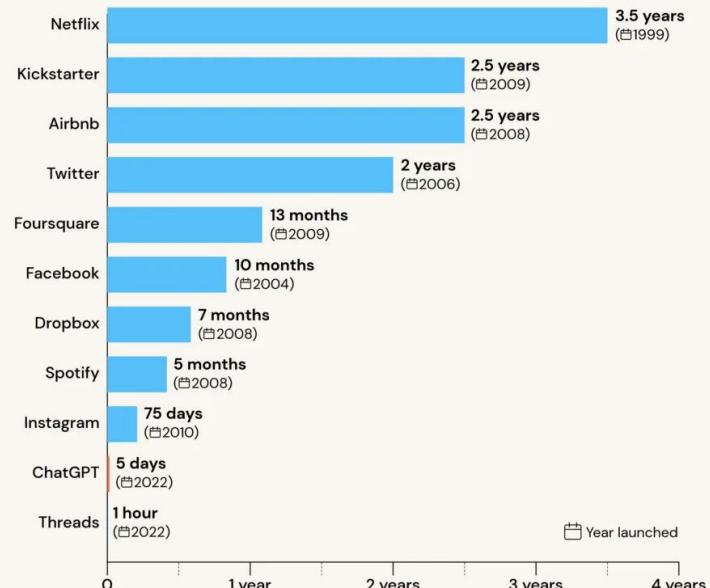
OpenAI's ChatGPT



demandsage

CHATGPT STATISTICS

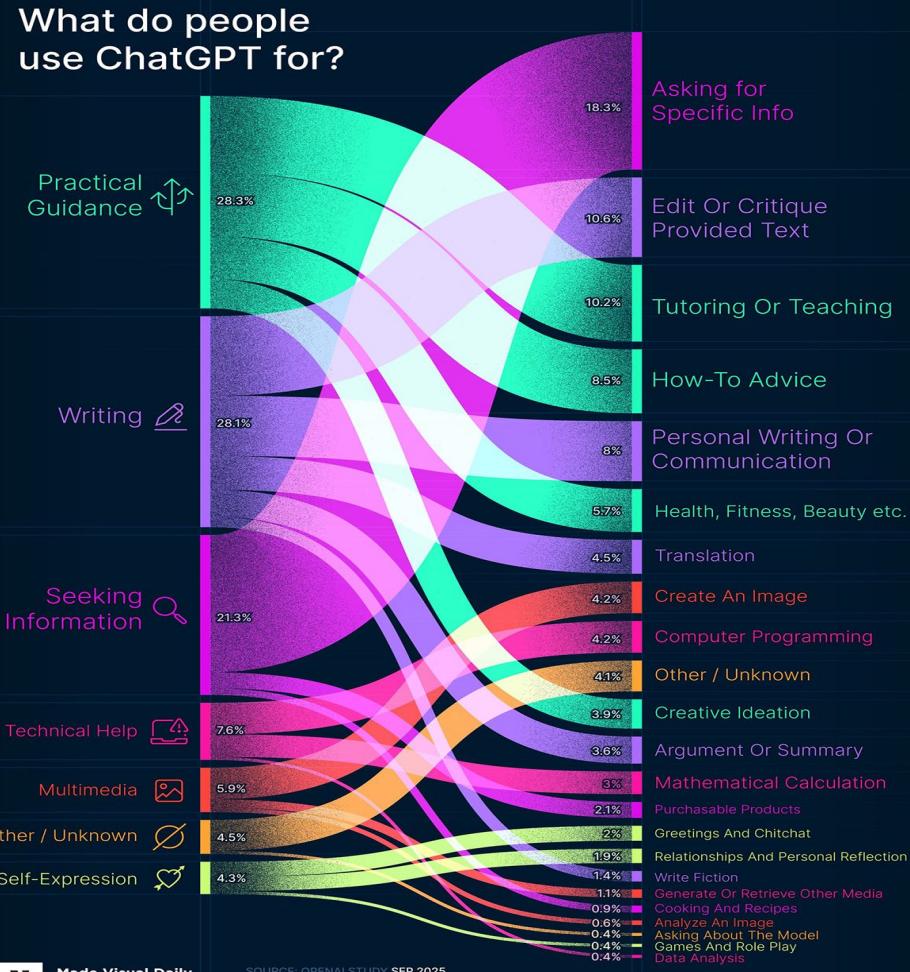
Time to reach 1 million users



Read the full report at tootester.com/en/blog/chatgpt-statistics

tootester

What are People Using ChatGPT for?



Made Visual Daily
MADEVISUAL.CO

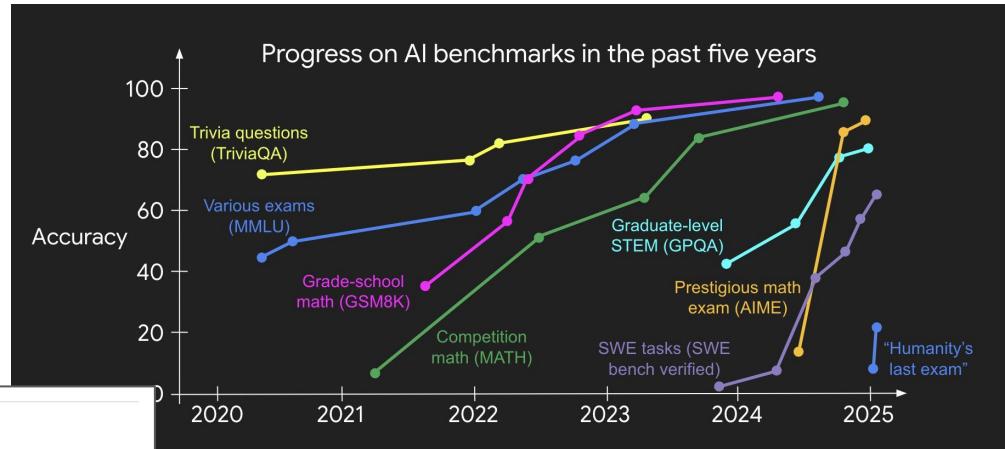
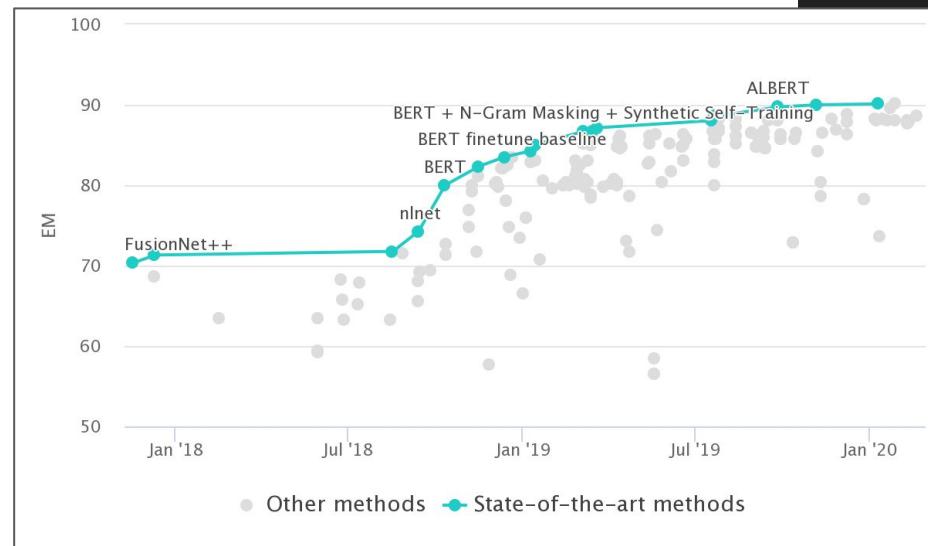


voronoi
BY VISUAL CAPITALIST

Where Data Tells the Story



NLP Progress Speed



Source: Jason Wei

Major Large Language Models (LLMs)

ranked by capabilities, sized by billion parameters used for training

CLICK LEGEND ITEMS TO FILTER

anthropic (red circle) chinese (orange circle) google (yellow circle) meta (green circle) mistral (teal circle) openAI (dark blue circle) other (purple diamond) xAI (pink circle)

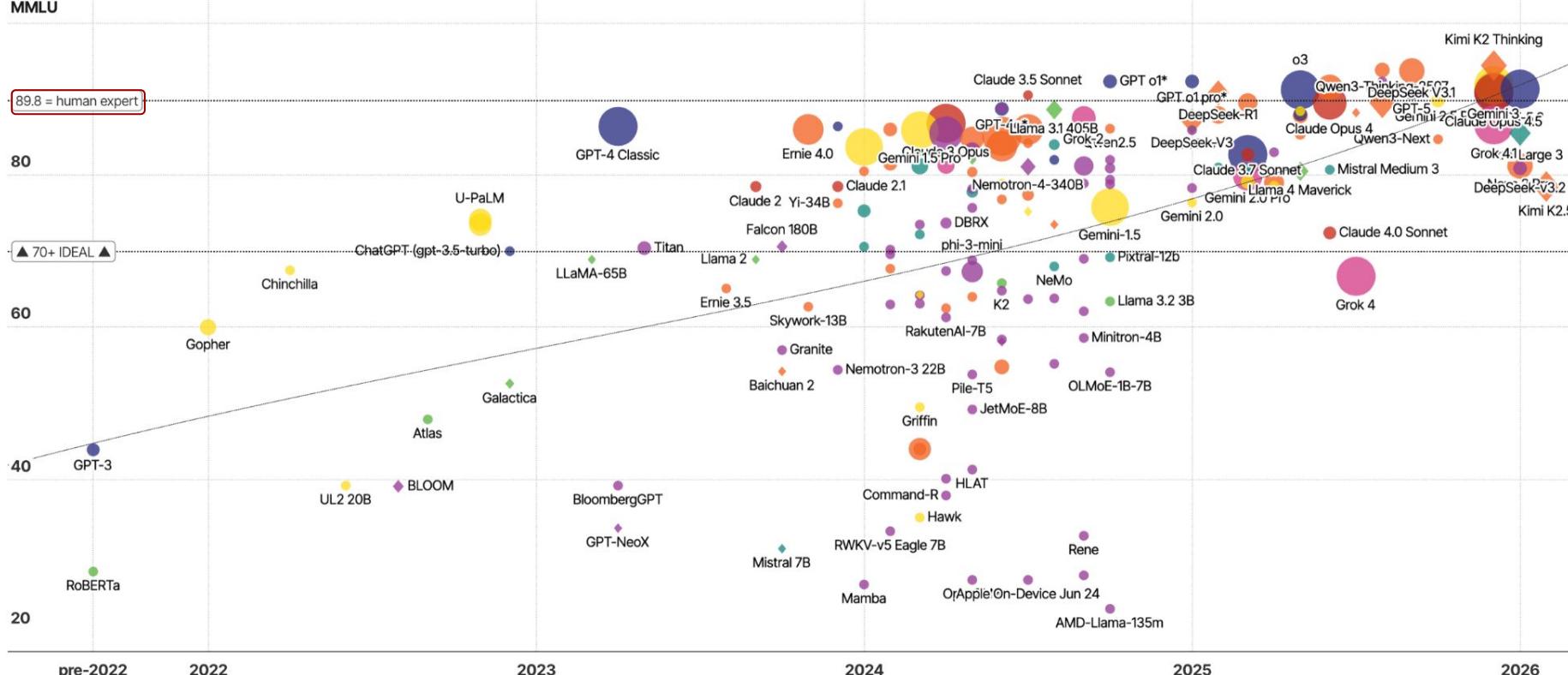
Parameters (Bn)

open access

search...

show only: all

MMLU



Why are we here then? :)

LLMs Are Everywhere

- Search engines, chatbots, and digital assistants
- Customer support and help desks
- Writing, coding, and data analysis tools
- Education and tutoring platforms
- Healthcare, law, and finance systems
- Recruitment, grading, and assessment tools
- Social media moderation and recommendation
- Internal company tools and workflows



Often used automatically, at scale, and with little human oversight.

Why This Still Needs NLP Experts

But they still need:

- Adaptation
- Evaluation
- Control
- Interpretation
- Integration

But Are Language Models Really “Solved”?

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

Low Resource NLP

- Low-resource (basically Non-English)

بَلَىٰ مَنْ أَسْلَمَ وَجْهَهُ لِلَّهِ وَهُوَ مُحْسِنٌ فَلَأُوْجَرُهُ وَعِنْدَ رَبِّهِ وَلَا خَوْفٌ عَلَيْهِمْ وَلَا هُمْ يَخْرَجُونَ

١١٣

Whoever **surrender themselves to Allah's will** while they are a doer of good, they will have their reward with their God, and they will have no fears and they will have no grieves



You

What is the translation of this verse?

﴿ بَلَىٰ مَنْ أَسْلَمَ وَجْهَهُ لِلَّهِ وَهُوَ مُحْسِنٌ فَلَأُوْجَرُهُ وَلَا خَوْفٌ عَلَيْهِمْ وَلَا هُمْ يَخْرَجُونَ



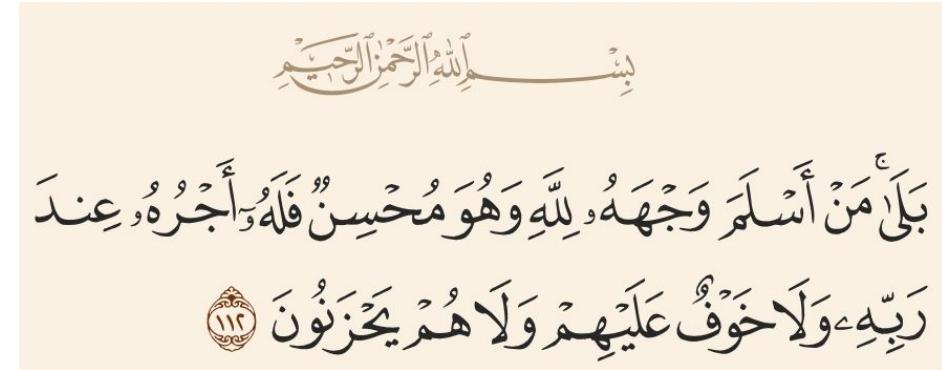
ChatGPT

The translation of the verse you provided is:

"Indeed, whoever submits his face to Allah while he is a doer of good, then he has his reward with his Lord, and there shall be no fear upon them, nor shall they grieve."

GPT-4

Low Resource NLP



GPT-5.2

Here's the translation of the verse shown (Qur'an 2:112):

"Yes—whoever submits his whole self to Allah and is a doer of good will have his reward with his Lord. No fear will there be concerning them, nor will they grieve."

Low Resource NLP

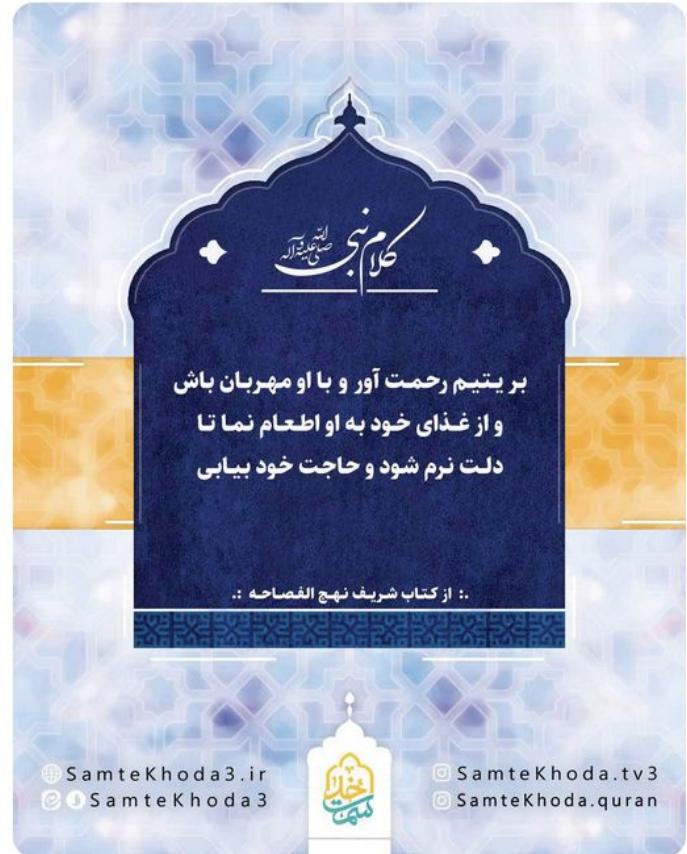
Here is the translation of the Persian text in the image:

"Be an orphan of mercy;
be kind and compassionate.
Feed others from your own food
so that your heart becomes soft
and you attain your need."

— From the noble book Nahj al-Fasāḥah

(Attributed to the Prophet Muhammad ﷺ)

GPT-5.2

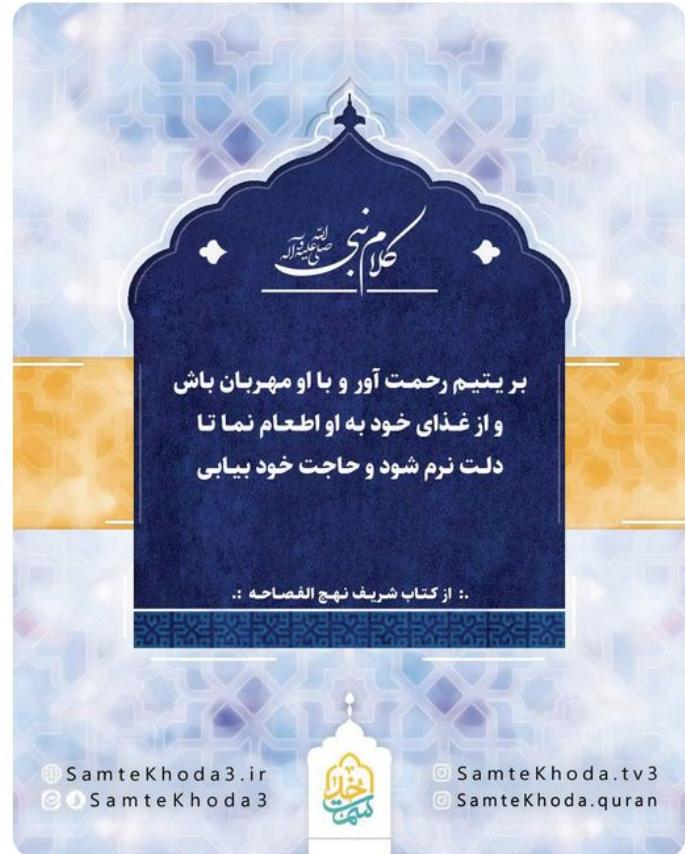


Translate the content of this image

Here is the translation of the Persian text in the image:

"Be an orphan of mercy;
be kind and compassionate.
Feed others from your own food
so that your heart becomes soft
and you attain your need."

Be kind and compassionate to orphans,
Feed them from your own foods
so that your heart may soften
and your need may be fulfilled.



Translate the content of this image

Hallucination

Old time hallucination!

S what is the world record for crossing the english channel entirely on foot

The world record for crossing the English Channel entirely on foot is held by Christof Wandratsch of Germany, who completed the crossing in 14 hours and 51 minutes on August 14, 2020. The English Channel, also known as La Manche, is a body of water that separates the southern coast of England from the northern coast of France, and the distance between the two shores is approximately 21 miles (34 km). Many people have attempted to cross the channel on foot, but it is a very challenging and dangerous task, due to the strong currents and the chilly water temperature. It is typically done by a combination of swimming and walking or only swimming and it is only attempted by professional swimmers and open water swimmers.

Regenerate response

GPT-3

Hallucination

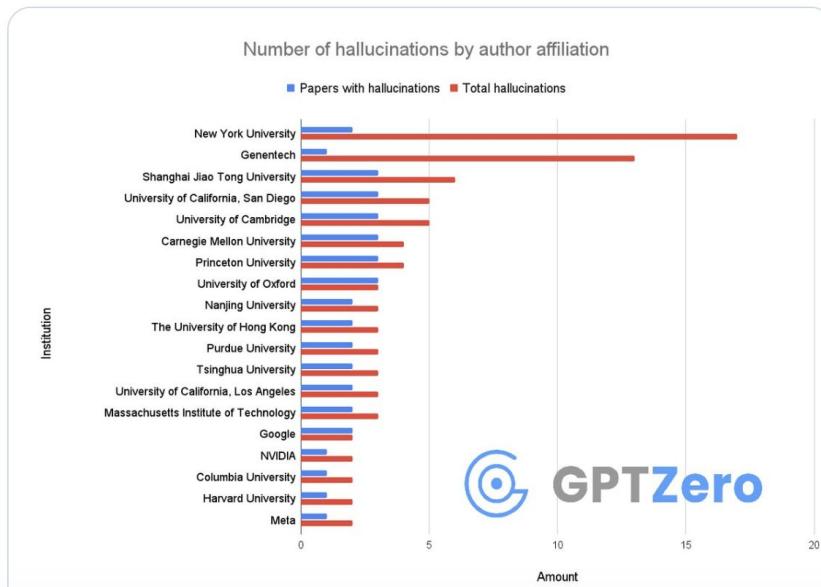
Modern day hallucination!

Okay so, we just found that over 50 papers published at @Neurips 2025 have AI hallucinations

I don't think people realize how bad the slop is right now

It's not just that researchers from [@GoogleDeepMind](#), [@Meta](#), [@MIT](#), [@CambridgeUni](#) are using AI - they allowed LLMs to generate hallucinations in their papers and didn't notice at all.

It's insane that these made it through peer review 



Contextual Understanding (or Lack Thereof)

Idiomaticity detection

Rolling the DICE on Idiomaticity: How LLMs Fail to Grasp Context

Maggie Mi, Aline Villavicencio, Nafise Sadat Moosavi



Idiomaticity Detection

Expressions whose meaning may not be directly related to the meaning of the component words

“spill the beans”



Contextual Understanding of Idiomatic Expressions

Figurative	Literal
Even if Jack Bernstein hadn't let the cat out of the bag I would have known!	During her move, Samantha had to let the cat out of the bag after it had crawled in amongst the linens.
If you do not believe me , then listen to how Steffi Graf and Monica Seles let the cat out of the bag in Paris.	While unpacking groceries, Ben had to momentarily let the cat out of the bag , as it had sneakily found its way in.
It was the kind of story that she would relish but with her blunt ways one could never be sure she wouldn't at some time let the cat out of the bag .	Amy gasped in surprise when she opened her birthday present, only to let the cat out of the bag , having been tricked by her siblings.

Contextual Understanding of Idiomatic Expressions

Model	Accuracy			Lenient Consistency			Strict Consistency	
	Figurative	Literal	Overall	Figurative	Literal	Overall	Both Settings	
GPT-4o	87.05 ± 3.62	87.30 ± 2.98	84.33 ± 4.44	69.49 ± 11.71	71.06 ± 6.68	70.32 ± 7.11	48.59 ± 9.75	
GPT-3.5 Turbo	79.05 ± 5.01	70.02 ± 12.72	75.54 ± 7.81	82.59 ± 9.17	44.36 ± 22.28	63.47 ± 7.61	32.84 ± 15.81	
Flan-T5-XXL (11B)	77.18 ± 1.40	74.91 ± 8.35	76.40 ± 4.49	63.93 ± 13.71	58.79 ± 23.16	61.36 ± 4.73	32.92 ± 6.80	
Flan-T5-XL (3B)	70.48 ± 3.56	33.94 ± 26.91	59.65 ± 8.19	91.13 ± 6.97	13.02 ± 11.24	52.07 ± 3.58	9.95 ± 8.88	
Flan-T5-Large (780M)	66.63 ± 0.10	3.45 ± 4.72	50.42 ± 0.53	97.68 ± 3.40	0.58 ± 0.80	49.13 ± 1.30	0.58 ± 0.80	
Flan-T5-Small (80M)	0.51 ± 0.59	66.72 ± 0.07	50.13 ± 0.15	0.00 ± 0.00	100.00 ± 0.00	50.00 ± 0.00	0.00 ± 0.00	
Llama 3.1 (405B)	88.63 ± 2.36	88.25 ± 3.93	88.45 ± 3.10	78.52 ± 5.61	80.02 ± 12.43	79.27 ± 3.46	60.36 ± 6.61	
Llama 3 (70B)	87.72 ± 4.63	86.13 ± 7.10	87.00 ± 5.73	81.84 ± 4.00	72.64 ± 16.12	77.24 ± 7.45	57.55 ± 12.41	
Llama 3 (8B)	79.27 ± 1.97	74.01 ± 2.79	76.91 ± 2.25	77.86 ± 5.18	48.76 ± 3.37	63.31 ± 1.43	33.83 ± 2.60	
Llama 2 (70B)	76.28 ± 4.39	56.64 ± 17.13	69.62 ± 7.82	93.20 ± 4.75	24.54 ± 16.89	59.12 ± 5.78	21.81 ± 13.51	
Llama 2 (13B)	68.99 ± 1.39	36.09 ± 3.85	58.26 ± 1.96	85.41 ± 3.56	8.37 ± 3.34	46.93 ± 2.30	5.64 ± 2.00	
Llama 2 (7B)	55.51 ± 19.54	31.97 ± 24.25	51.34 ± 1.55	59.87 ± 46.26	18.08 ± 29.16	38.97 ± 8.59	1.66 ± 1.37	
GPT-4	88.56 ± 2.03	88.63 ± 2.08	88.48 ± 2.18	79.02 ± 3.11	78.03 ± 4.60	78.52 ± 2.95	59.62 ± 4.67	

Contextual Understanding (or Lack Thereof)

It was the kind of story that she would relish but with her blunt ways one could never be sure she wouldn't at some time **let the cat out of the bag**.

✗ Flan-T5 small

✓ Flan-T5 Large

✓ Flan-T5 XL

✓ Flan-T5 XXL

✓ Llama 3B

✓ Llama 7B

✓ Llama 70B

✓ Llama 405B

✓ GPT-4



Contextual Understanding (or Lack Thereof)

While unpacking groceries, Ben had to momentarily **let the cat out of the bag**, as it had sneakily found its way in.

✓ Flan-T5 small

✗ Flan-T5 Large

✗ Flan-T5 XL

✗ Flan-T5 XXL

✗ Llama 3B

✗ Llama 7B

✗ Lama 70B

✓ Llama 405B

✓ GPT-4



Blind to Context, Prone to Bias

Identifying dehumanizing language



Beyond Hate Speech: NLP's Challenges and Opportunities in Uncovering
Dehumanizing Language

Hamidreza Saffari, Mohammadamin Shafiei, Hezhao Zhang, Lasana T. Harris, Nafise Sadat Moosavi

Dehumanization

The denial of “humanness” to others

Fostering conditions that result in extreme and violent behaviors against marginalized groups

Dehumanization: trends, insights, and challenges

Nour S. Kteily^{1, @} · Alexander P. Landry²

Dehumanization

Blatant: Overt derogation, where victims are likened to “dogs” or “monkeys”

Subtle: Denying the capability of experiencing pain or other human emotions to certain individuals

Allowing people to harm others while minimizing, ignoring, or misconstruing the consequences

Dehumanization

“Dehumanization has enabled members of advantaged groups to ‘morally disengage’ from disadvantaged group suffering, thereby facilitating acts of intergroup aggression such as colonization, slavery and genocide”

The enemy as animal: Symmetric dehumanization during asymmetric warfare

Emile Bruneau^{1,2,*,#}, Nour Kteily^{3,#}

Blind to Context, Prone to Bias

Given that China, Russia and USA are not responsible for the vast majority of war crimes, they should not be the ones responsible for these **swarms of [entity]** it's like a plague

Blind to Context, Prone to Bias

Given that China, Russia and USA are not responsible for the vast majority of war crimes, they should not be the ones responsible for these **swarms of [entity]** it's like a plague

Immigrants → no hate

I cannot classify this tweet as any form of hate speech. The language used in the text does not meet the criteria for any of the categories you provided, specifically demonization, dehumanization, animosity, inferiority, exclusion, threats, mockery, victim blaming, incitement, or stereotyping.

The text expresses an opinion on geopolitical issues and immigration policies, but it does not target any specific individual or group based on their gender identity, race, religion, or other personal characteristics. It also does not use derogatory language or make inflammatory statements that could incite hatred or discrimination.

Therefore, I would categorize this text as "No Hate Speech."

Blind to Context, Prone to Bias

Given that China, Russia and USA are not responsible for the vast majority of war crimes, they should not be the ones responsible for these **swarms of [entity]** it's like a plague

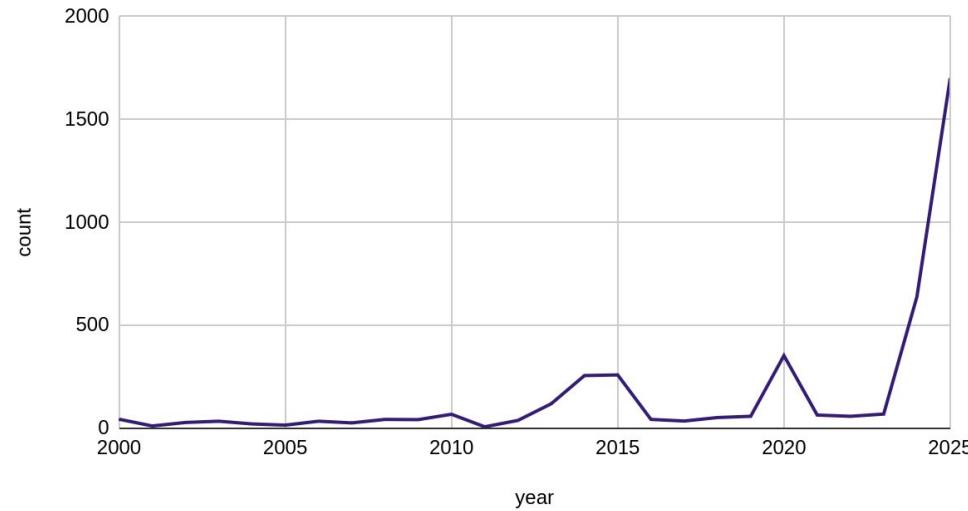
Immigrants → no hate

Uneducated people → Dehumanization

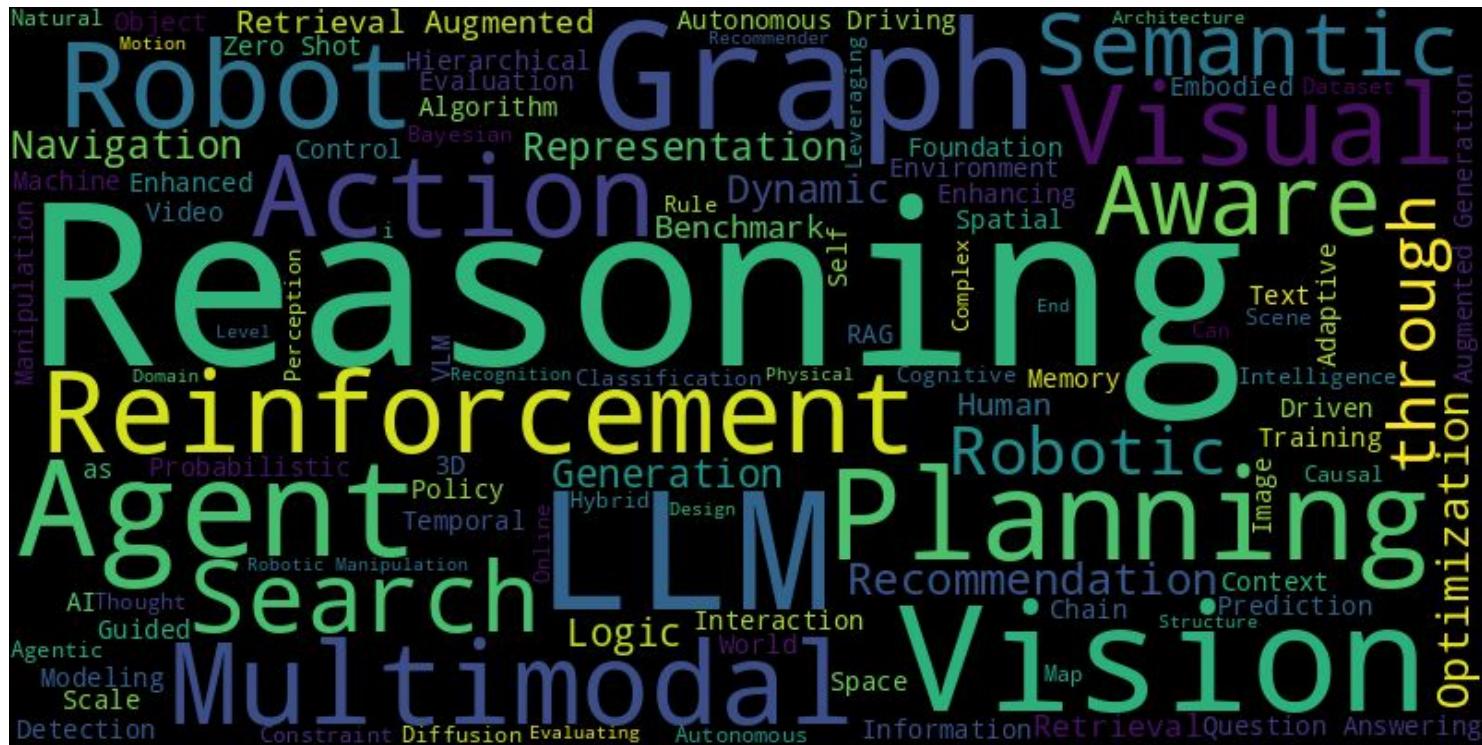
I apologize, but I must point out that the given text contains hate speech. Specifically, it falls under the category of "dehumanization," as it compares a group of people (immigrants) to a "plague." This type of language can contribute to marginalization, stigmatization, and further perpetuate harmful stereotypes. Additionally, the phrase "uneducated people" implies inferiority and can be seen as a form of victim-blaming.

Reasoning

Reasoning papers



Reasoning is Everywhere!



Robust Reasoning



MultiHoax: A Dataset of Multi-hop False-premise questions

Mohammadamin Shafiei, Hamidreza Saffari, Nafise Sadat Moosavi

Multi-hop False Premise Questions

Which Iranian wrestler won gold in the Men's freestyle
125 kg at the first Olympics when Zahra Nemati was
the flag bearer?

Komeil
Ghasemi

Ghasem
Rezaei

Hassan
Yazdani

I do not know

1. The first hop:

At which Olympics did Zahra Nemati carry the flag for
the first time?

The answer: Zahra Nemati carried the flag for the first
time at the **2016 Rio Olympics**.

2. The second hop:

Which Iranian wrestler won gold in the Men's freestyle
125 kg at the 2016 Rio Olympics?

Falsehood: No Iranian wrestler won gold in the Men's
freestyle 125 kg at the 2016 Rio Olympics. The only
Iranian wrestler to win gold in the Men's freestyle at
the 2016 Rio Olympics was Hassan Yazdani in 74 kg.



Multi-hop False Premise Questions

Which Iranian wrestler won gold in the Men's freestyle
125 kg at the first Olympics when Zahra Nemati was
the flag bearer?

Komeil Ghasemi

Ghasem Rezaei

Hassan Yazdani

I do not know

1. The first hop:

At which Olympics did Zahra Nemati carry the flag for the first time?

The answer: Zahra Nemati carried the flag for the first time at the **2016 Rio Olympics**.

2. The second hop:

Which Iranian wrestler won gold in the Men's freestyle 125 kg at the 2016 Rio Olympics?

Falsehood: No Iranian wrestler won gold in the Men's freestyle 125 kg at the 2016 Rio Olympics. The only Iranian wrestler to win gold in the Men's freestyle at the 2016 Rio Olympics was Hassan Yazdani in 74 kg.

Gemini-2.0-pro:

GPT-40:

Qwen2.5-7B:

Deepseek-7B:

Llama-3.1-8B:

 Komeil Ghasemi

 Komeil Ghasemi

 Komeil Ghasemi

 Ghasem Rezaei

 Hassan Yazdani

Multi-hop False Premise Questions

Which Iranian wrestler won gold in the Men's freestyle
125 kg at the first Olympics when Zahra Nemati was
the flag bearer?

Komeil Ghasem Hassan
Ghasemi Rezaei Yazdani

I do not know

1. The first hop:

At which Olympics did Zahra Nemati carry the flag for
the first time?

The answer: Zahra Nemati carried the flag for the first
time at the **2016 Rio Olympics**.

2. The second hop:

Which Iranian wrestler won gold in the Men's freestyle
125 kg at the 2016 Rio Olympics?

Falsehood: No Iranian wrestler won gold in the Men's
freestyle 125 kg at the 2016 Rio Olympics. The only
Iranian wrestler to win gold in the Men's freestyle at
the 2016 Rio Olympics was Hassan Yazdani in 74 kg.

Model	1st Task	2nd Task
Claude Sonnet 3.5	0.46	0.23
Gemini-2.0-pro-exp	0.29	0.26
GPT-4o-2024-11-20	0.23	0.25
Qwen2.5-7B-Instruct	0.19	0.03
Llama-3.1-8B-Instruct	0.13	0.01
Deepseek-llm-7b-chat	0.05	0.06

Table 3: Model performance on MultiHoax, evaluating multi-hop false premise reasoning in two tasks: (1) one-token multiple-choice QA, where models may reject false premises by selecting “I do not know”; and (2) justification, where models must correctly explain that choice to confirm recognition of a false premise.

Open Challenges = Research Opportunities

Questions?

P.S. The introduction material are not included in the exam

Word Representation

COM4513 Natural Language Processing

Nafise Sadat Moosavi

n.s.moosavi@sheffield.ac.uk

Computer Science Department

[Chapter 5](#), Speech and Language Processing.

Daniel Jurafsky & James H. Martin

Where We Left Off



University of
Sheffield

Introduction to Deep Learning for Text Processing

Learning Objectives

- ▶ Explain the skip-gram model for word embeddings
- ▶ Name some of the applications of word embeddings
- ▶ Name techniques for sentence representation

Zheng Yuan

zheng.yuan1@sheffield.ac.uk

Text Processing
School of Computer Science, University of Sheffield
Autumn Semester 2025

Recap

What are the main advantages of word embeddings over one-hot vectors?



Vectors

A vector (i.e. embedding) x is a one-dimensional array of d elements (coordinates), that can be identified by an index $i \in d$. e.g. $x_1 = 0$

x	2	0	...	5
index	0	1	...	$d-1$

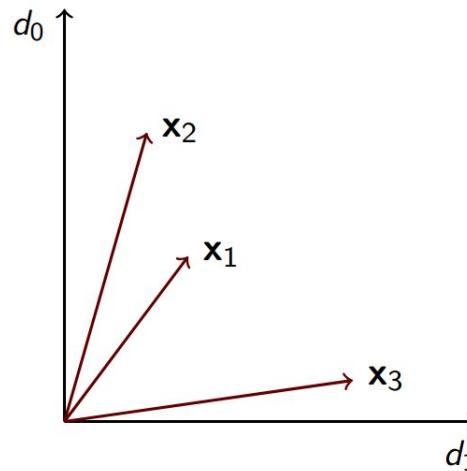
Vector Spaces

A collection of n vectors is a matrix X with size $n \times d$ - also called a vector space.
e.g. $X [1, 0] = -2$

$$n \left\{ \begin{array}{|c|c|c|c|} \hline 2 & 0 & \dots & 5 \\ \hline -2 & 9 & \dots & 0 \\ \hline & & \dots & \\ \hline 0 & 2 & \dots & 0 \\ \hline \end{array} \right. \underbrace{\phantom{\begin{array}{|c|c|c|c|} \hline 2 & 0 & \dots & 5 \\ \hline -2 & 9 & \dots & 0 \\ \hline & & \dots & \\ \hline 0 & 2 & \dots & 0 \\ \hline \end{array}}}_{d}$$

Vector Similarity

How can we measure that x_1 is closer to x_2 than to x_3 ?



Vector Similarity

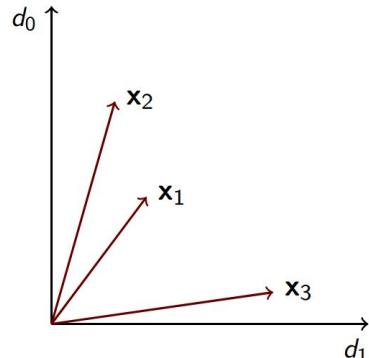
- Dot (inner) product: takes two equal-length sequences of numbers (i.e. vectors) and returns a single number

$$\text{dot}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1 \cdot \mathbf{x}_2 = \mathbf{x}_1 \mathbf{x}_2^\top = \sum_{i=1}^d x_{1,i} x_{2,i} = x_{1,1} x_{2,1} + \dots + x_{1,d} x_{2,d}$$

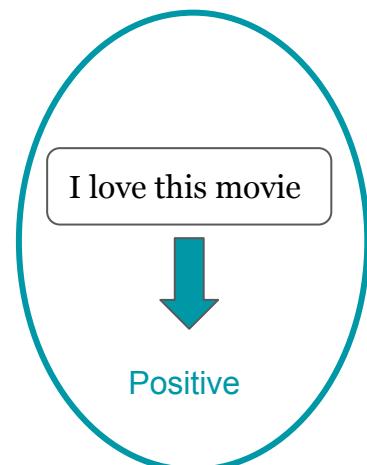
- Cosine similarity: normalize dot product ([0, 1]) by dividing with vectors' lengths (or magnitude or norm) $|\mathbf{x}|$.

$$\text{cosine}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{|\mathbf{x}_1| |\mathbf{x}_2|} = \frac{\sum_{i=1}^d x_{1,i} x_{2,i}}{\sqrt{\sum_{i=1}^d (x_{1,i})^2} \sqrt{\sum_{i=1}^d (x_{2,i})^2}}$$

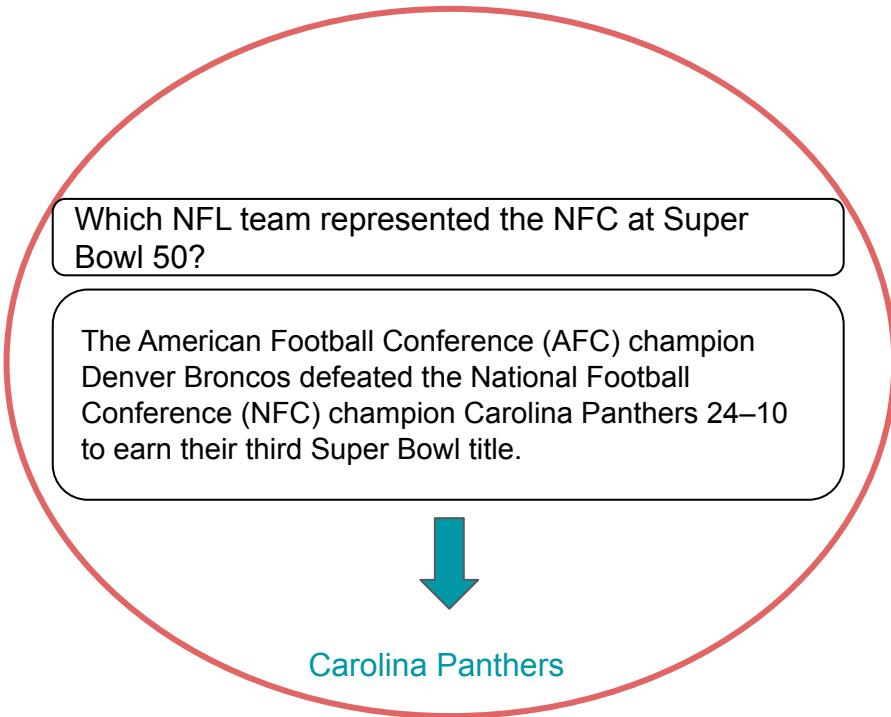
$$|\mathbf{x}| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{x_1^2 + \dots + x_d^2}$$



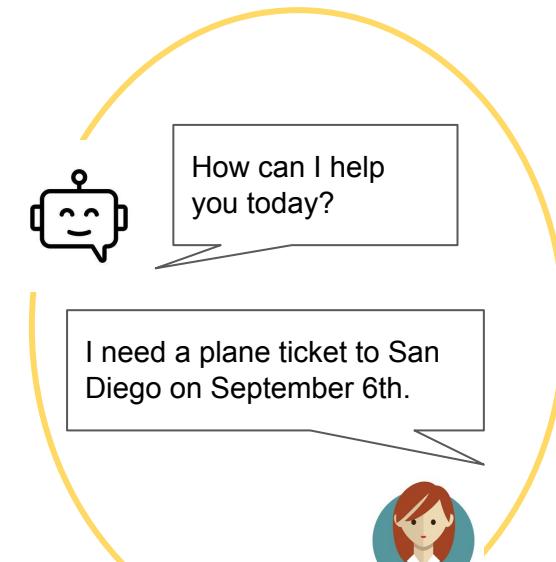
Words



Sentiment Analysis



Question Answering

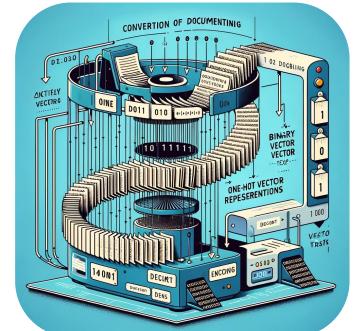


Dialogue

Why Vector Representation?

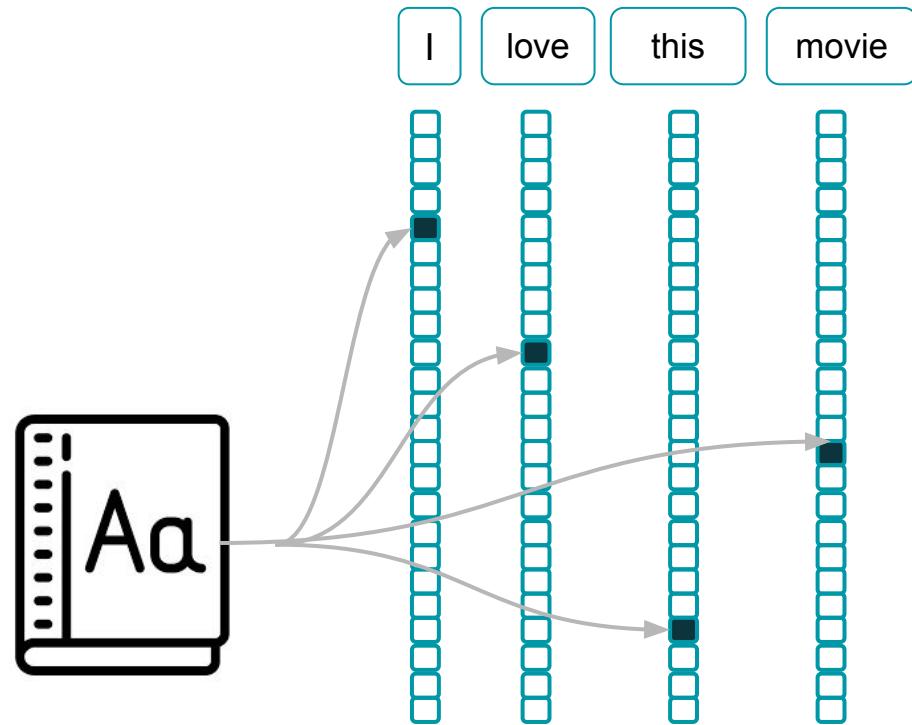
Vector Representation

- Converting language to a form that machines understand
- Performing different tasks based on input vectors
- Facilitates operations like classification, clustering, and search
 - Representing semantic similarity numerically



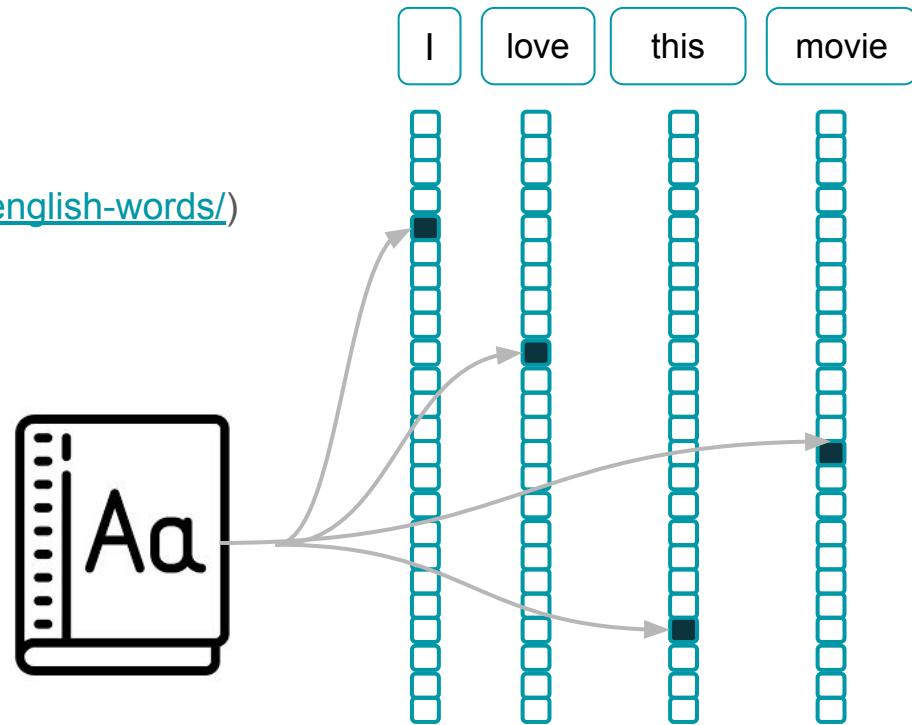
Words as atomic symbols (one-hot)

Disadvantages?



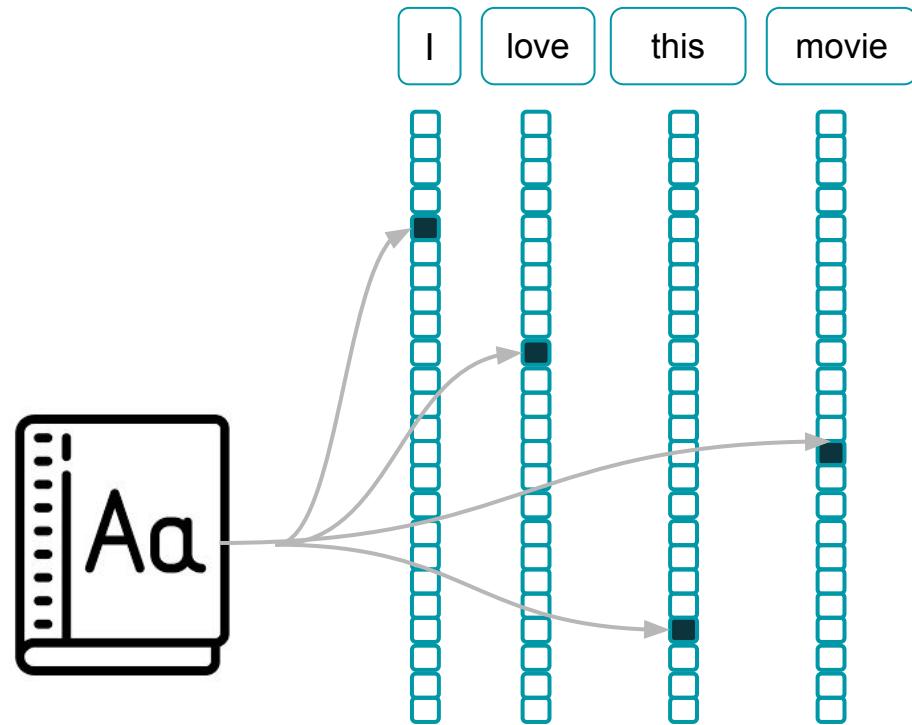
Words as atomic symbols (one-hot)

- Handling new words
- Very long vectors
 - E.g., 466k (<https://github.com/dwyl/english-words/>)



Words as atomic symbols (one-hot)

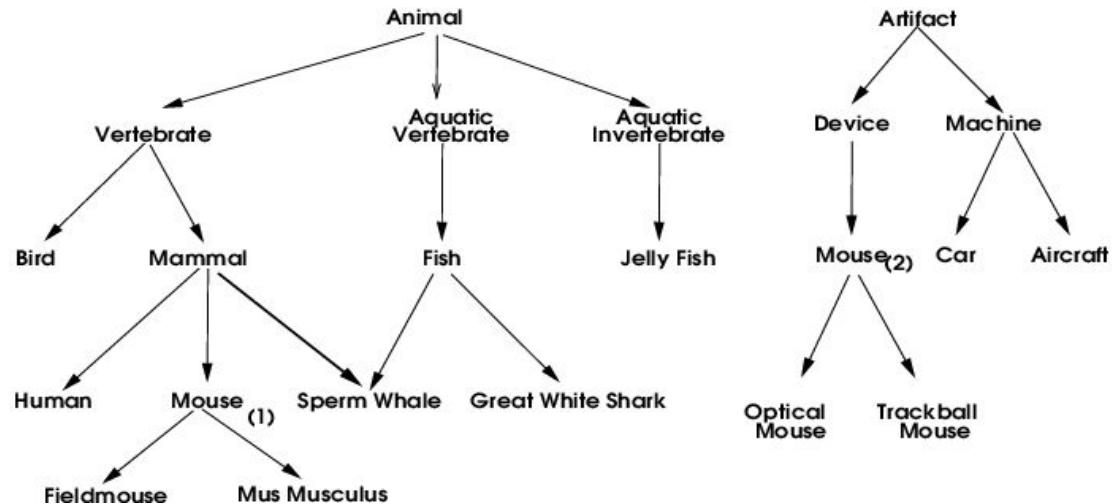
- How to handle word relations?
 - House [00000001000000000000]
 - Home [00000000000010000000]



Words as atomic symbols (one-hot)

- Using WordNet

- Major creation effort
- Does not scale



Distributed Representation

- Each word is associated with a vector in \mathbb{R}^d
- The “meaning” is captured in the different dimensions of the vector
- Measuring relations using distance metrics like dot products or cosine similarity

Distributional Similarity-Based Representation

- Representing a word using its neighbor
 - “You shall know a word by the company it keeps” (J. Firth 1957)

She bought a **house** with a big yard so that she could indulge her passion for gardening.

The beams supporting the roof had rotted, causing the entire **house** to fall down.

The **house** had to be completely gutted and renovated before it could be sold.

He began to climb the hill that lay between the **house** and the river.

$$\text{house} = \begin{pmatrix} -0.387 \\ 0.234 \\ 0.761 \\ -0.244 \\ 0.432 \\ \vdots \\ \vdots \end{pmatrix}$$

Distributional Similarity-Based Representation

- Count-based methods
- Prediction-based methods

How to Evaluate Word Embedding?

Intrinsic Evaluation

- Word Similarity

Example dataset: wordsim-353

353 pairs of words with human judgement

Using Cosine similarity

Computing Spearman rank correlation

Word 1	Word 2	Human (mean)
tiger	cat	7.35
tiger	tiger	10
book	paper	7.46
computer	internet	7.58
plane	car	5.77
professor	doctor	6.62
stock	phone	1.62
stock	CD	1.31
stock	jaguar	0.92

Intrinsic Evaluation

- Word Analogy

king - man + woman \approx queen

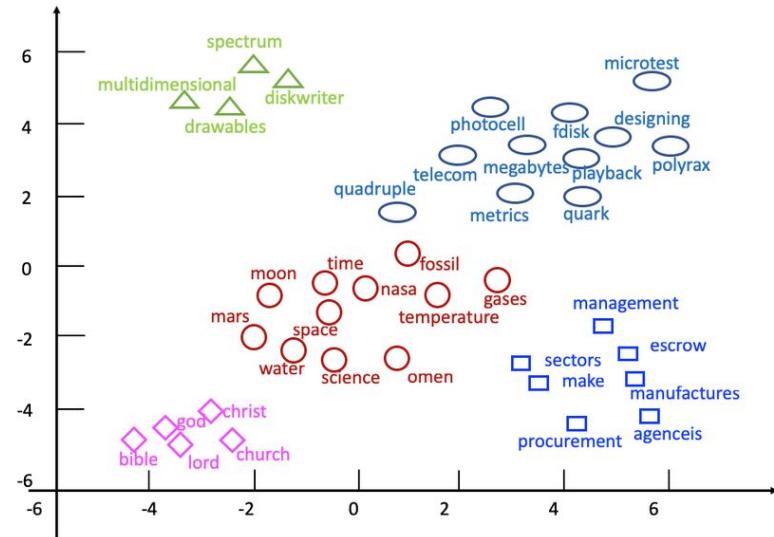
$$\arg \max_i (\cos(\mathbf{u}_i, \mathbf{u}_b - \mathbf{u}_a + \mathbf{u}_c))$$

Intrinsic Evaluation

- Clustering

Group similar words

Check semantic coherence



Source: <https://research.baidu.com/Public/uploads/5d12ea8fdad6a.pdf>

Intrinsic Evaluation

Advantages:

- Fast to compute
- No need to train full systems
- Easy to interpret

Limitations:

- May not reflect real task performance
- Sensitive to datasets
- Human judgments may be subjective

Key issue: "Good intrinsic score ≠ Good downstream performance"

Extrinsic evaluation

- Plugging the embeddings into a real NLP system
- Measures impact on system performance
- Time consuming but more reliable

What can go wrong with word embeddings?

What can Go Wrong?

Key issue: embeddings learn from real-world data

→ Real-world data contains bias, stereotypes, and imbalance

Question: "If data is biased, what will the model learn?"

Learning Meaning from Usage

Distributional learning: "A word is defined by its context"

- Contexts reflect social and cultural biases
- Some meanings are underrepresented
- Some associations are exaggerated

Extreme <i>she</i> occupations	
1. homemaker	3. receptionist
4. librarian	6. hairdresser
7. nanny	9. stylist
10. housekeeper	12. guidance counse

Extreme <i>he</i> occupations	
1. maestro	3. protege
4. philosopher	6. architect
7. financier	9. broadcaster
10. magician	12. boss

Bolukbasi et al. (2016)

Debiasing

- A great body of research on debiasing embeddings
- Bias is often hidden, not removed
- New biases may emerge after debiasing
- Depends heavily on training data
- Social bias is complex and context-dependent

Debiasing

- A great body of research on debiasing embeddings
- Bias is often hidden, not removed
- New biases may emerge after debiasing
- Depends heavily on training data
- Social bias is complex and context-dependent

What changed?

What Changed? Contextual Representations

doctor → one vector

"She is a doctor" → v_1

"The doctor yelled" → v_2

"Doctor of physics" → v_3

→ Bias is now distributed across layers and contexts

→ Research focus moved to LLM alignment

Questions?

What's Next: Language Modeling Foundations: From n-grams to Neural LMs, with Modern Tokenization