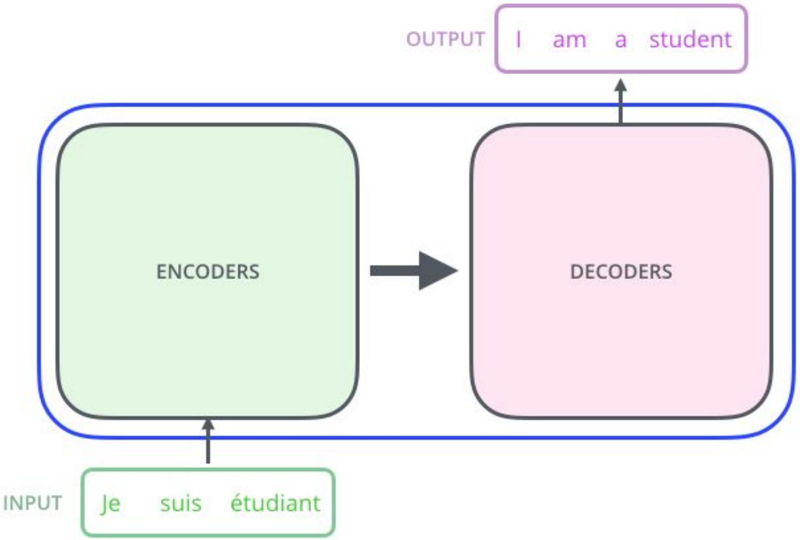
Large Language Models for Digital Humanities

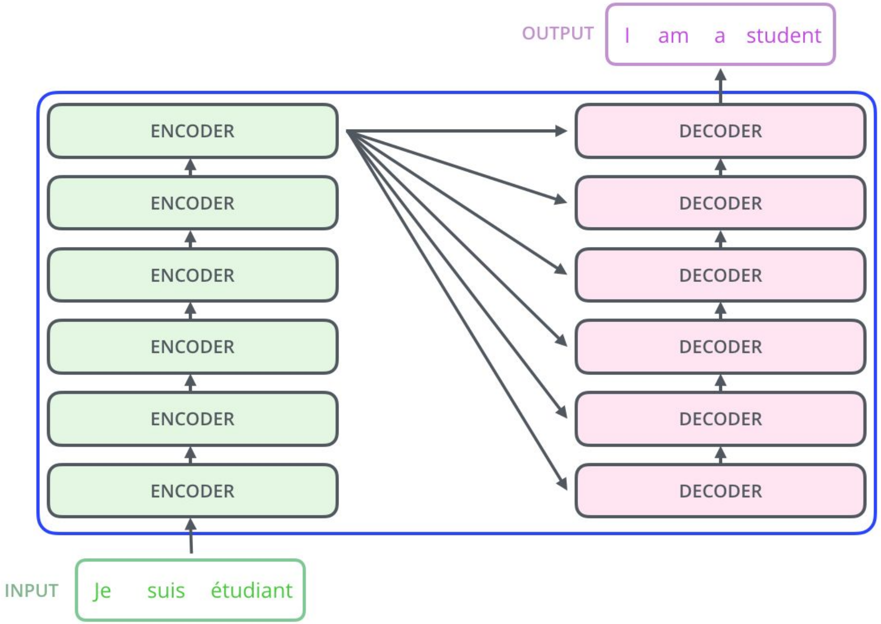
Animesh Mukherjee

Dept. of Computer Science and Engineering
IIT Kharagpur

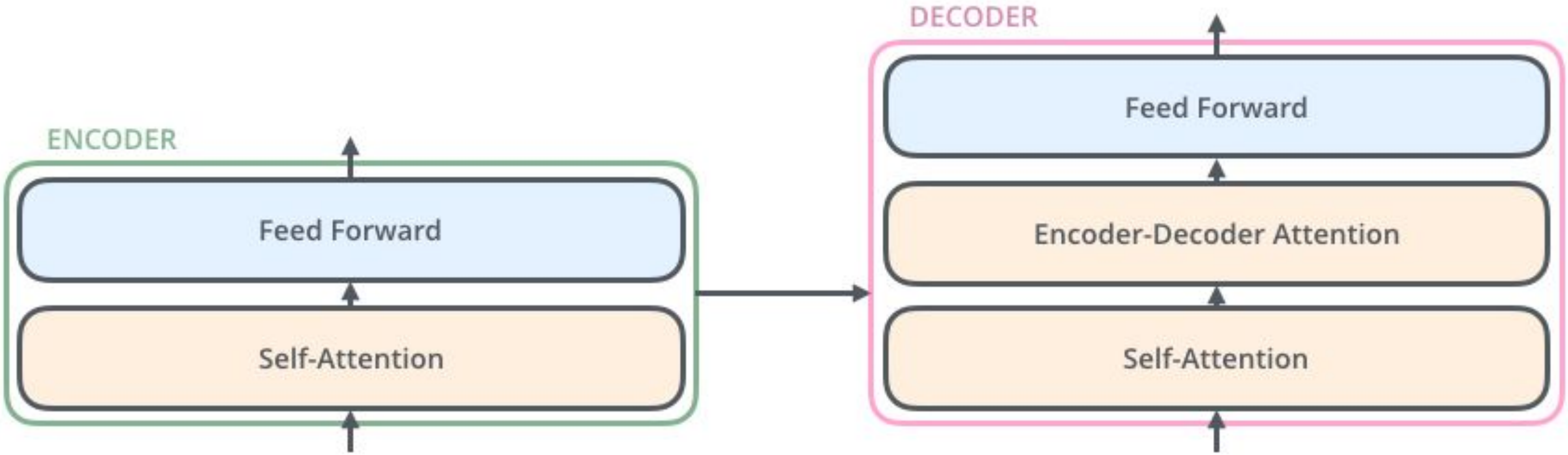
Transformers – The building blocks of LLMs

Teasing apart the transformer architecture

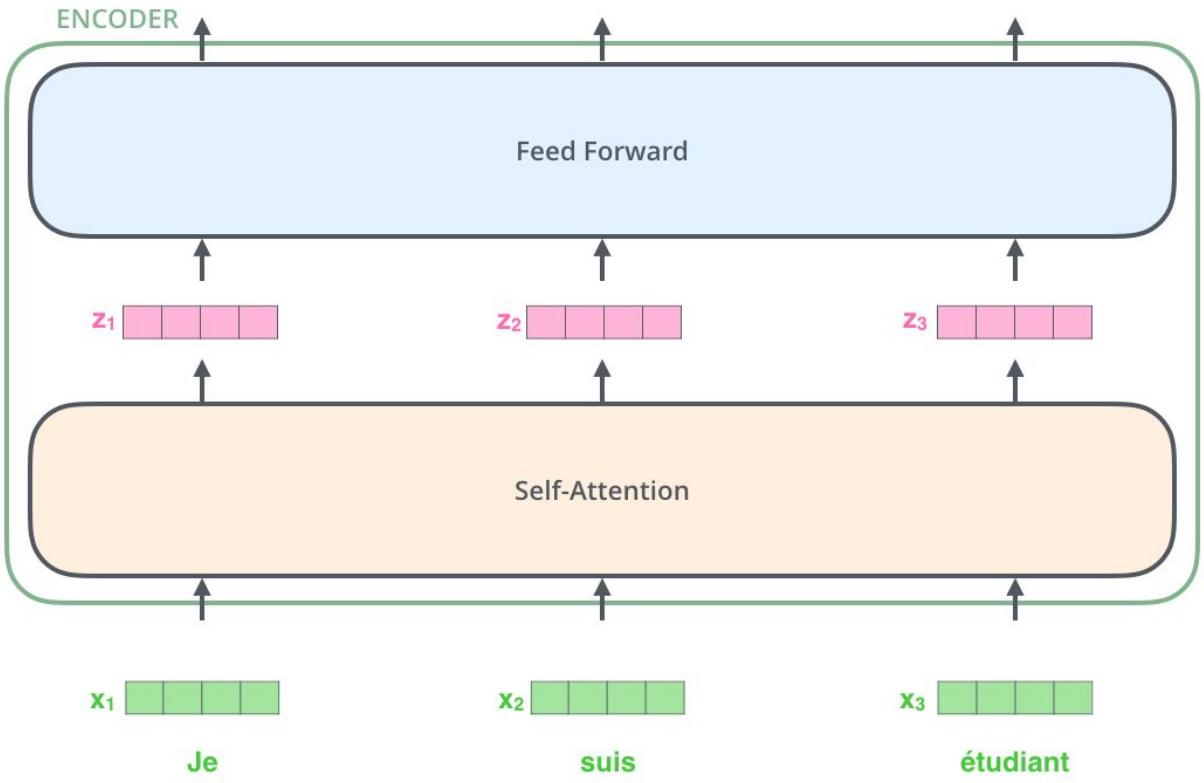




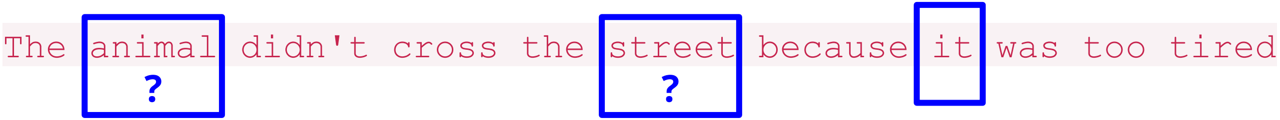
Transformer encoder and decoders



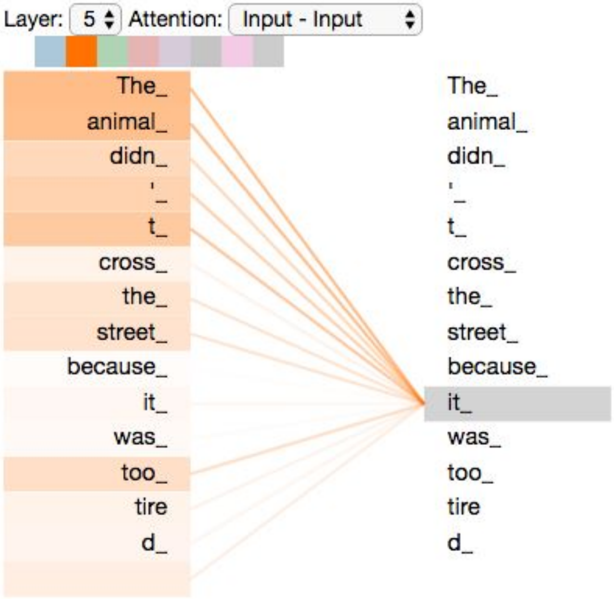
Encoder embeddings



Idea of self attention

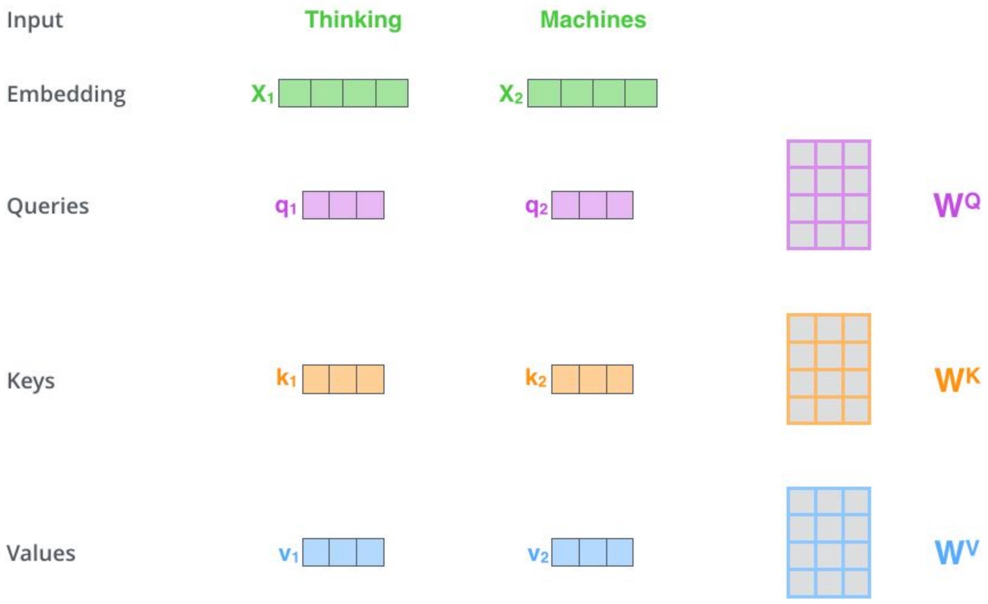


- When the model is processing the word “it”, self-attention allows it to associate “it” with “animal”.
- As the model processes each word (each position in the input sequence)
  - self attention allows it to look at other positions in the input sequence for clues
  - help lead to a better encoding for this word.

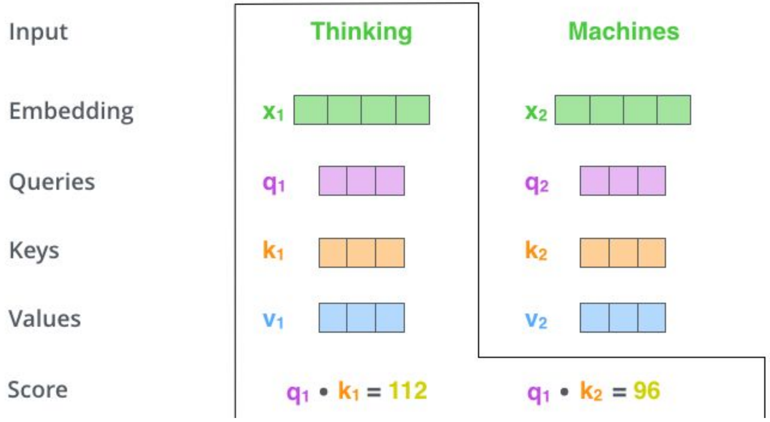


Self-attention calculation

Create three vectors - query, key and value for each word

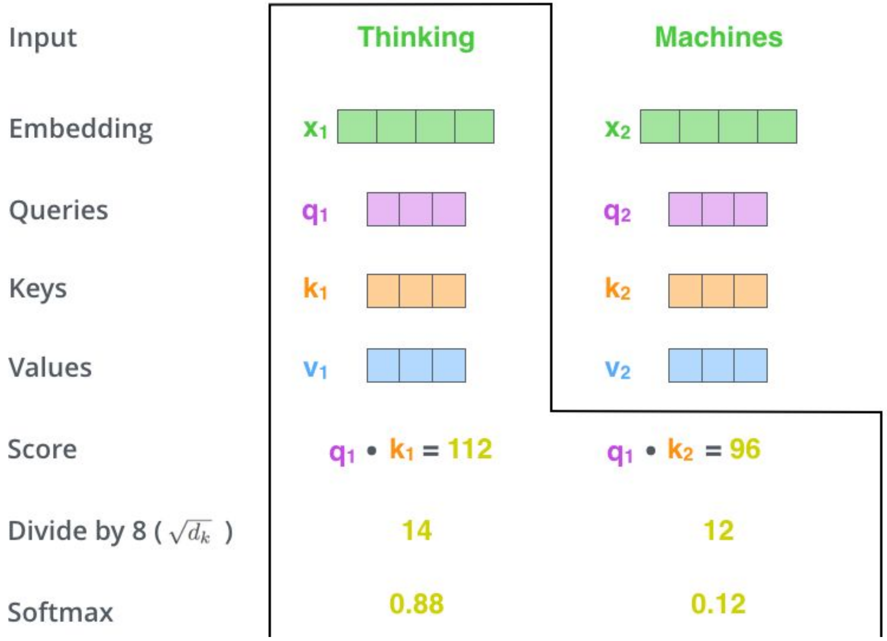


Attention score - dot product of query & key vectors

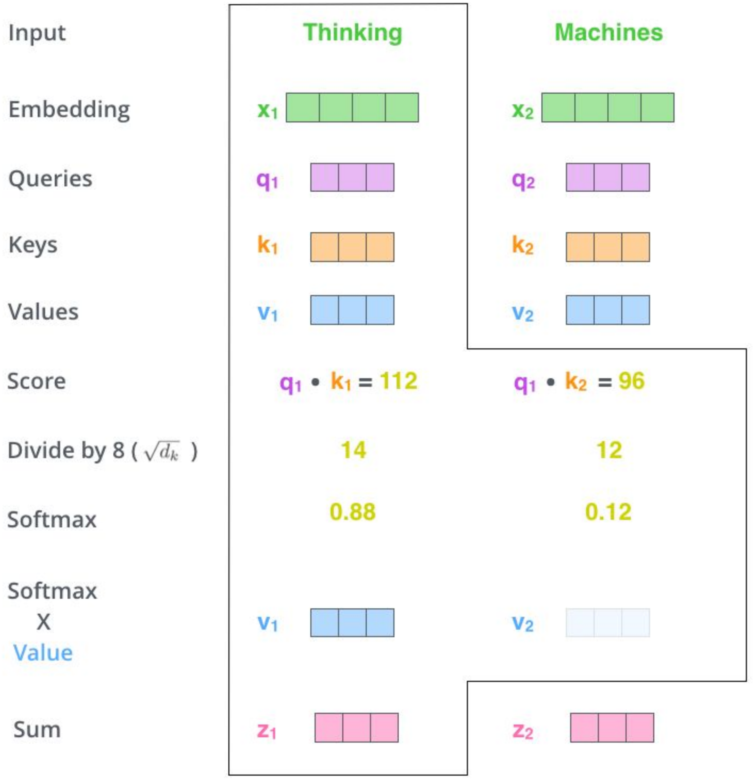


Self-attention calculation

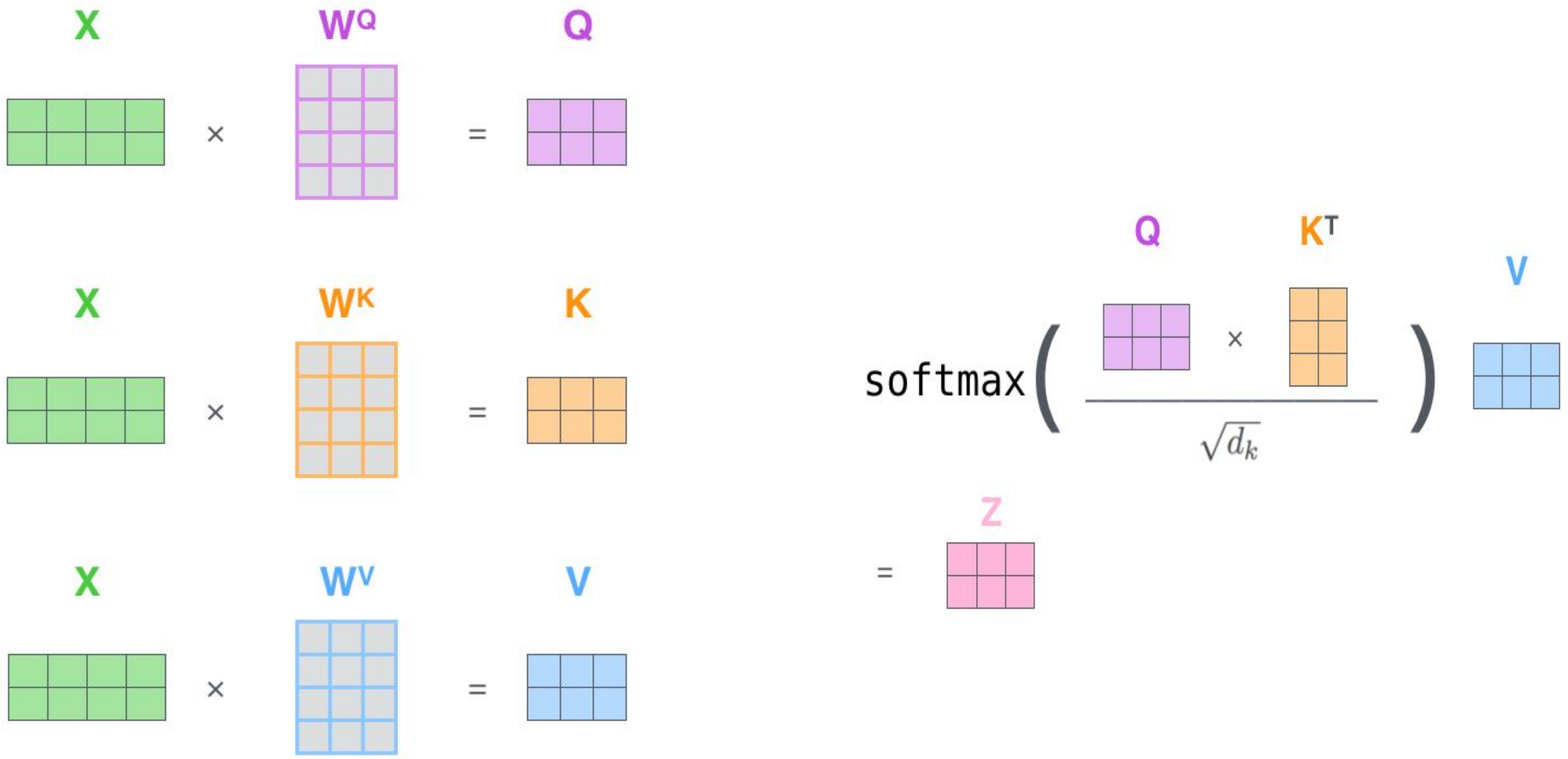
Normalize and softmax the attention score



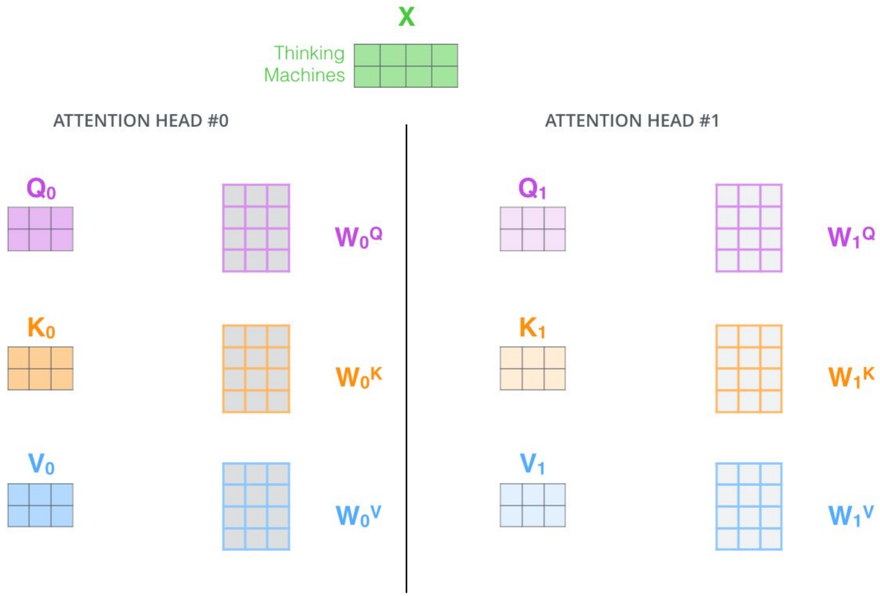
Attention score X softmax, sum up the value vectors at each word position

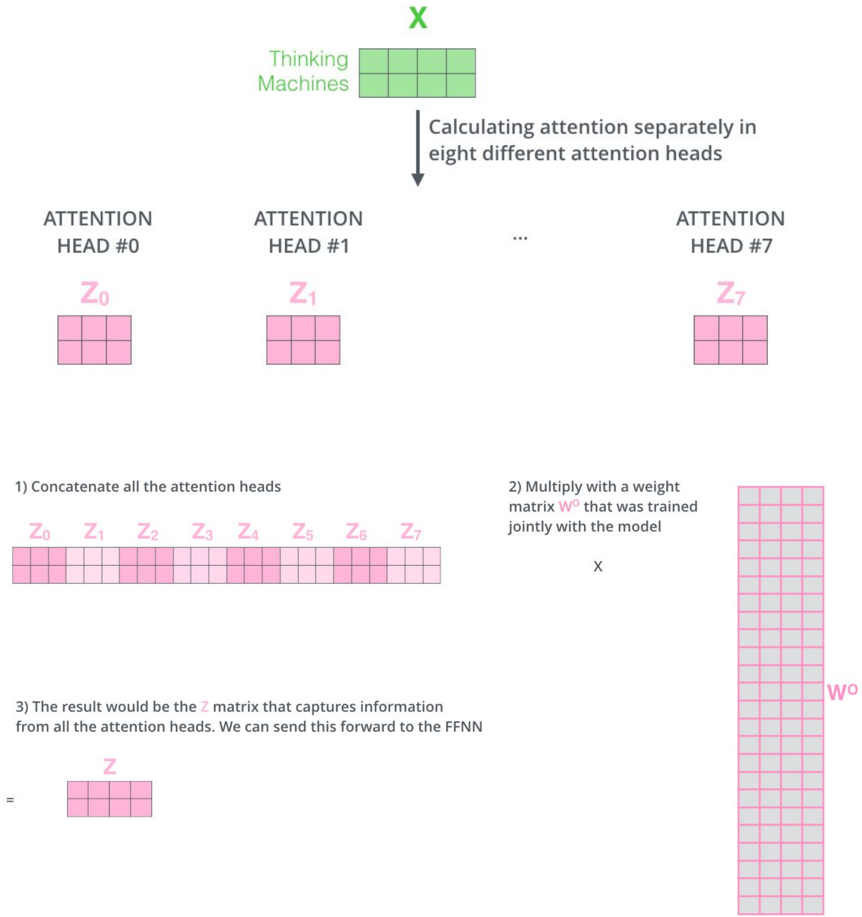


Calculation using matrices

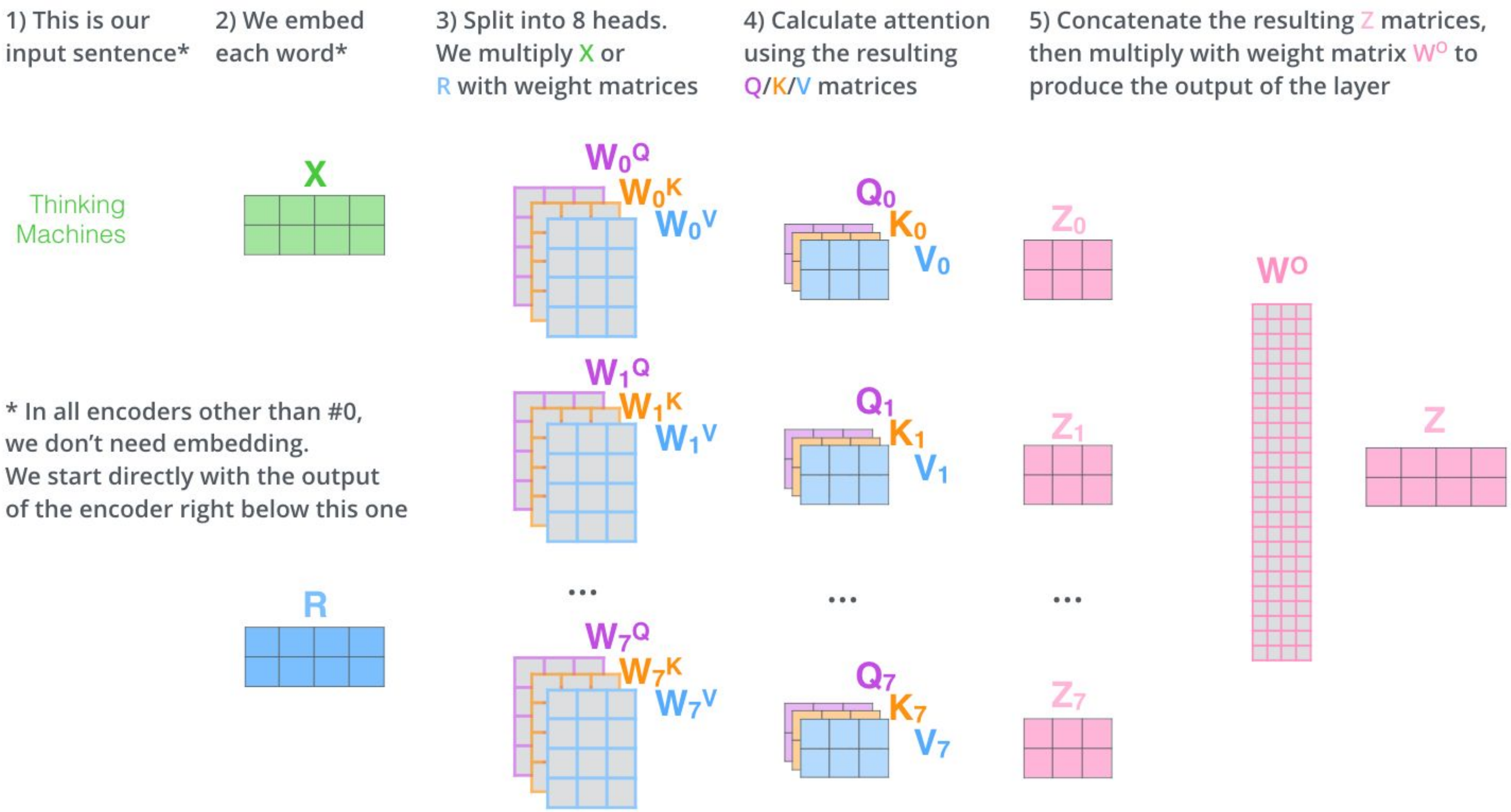


Multi-head attention

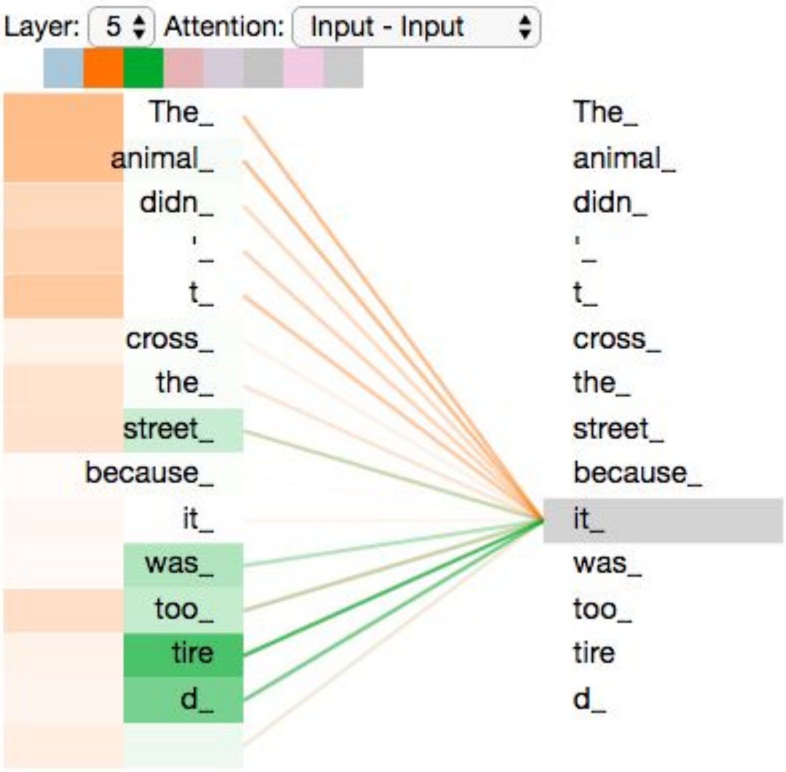




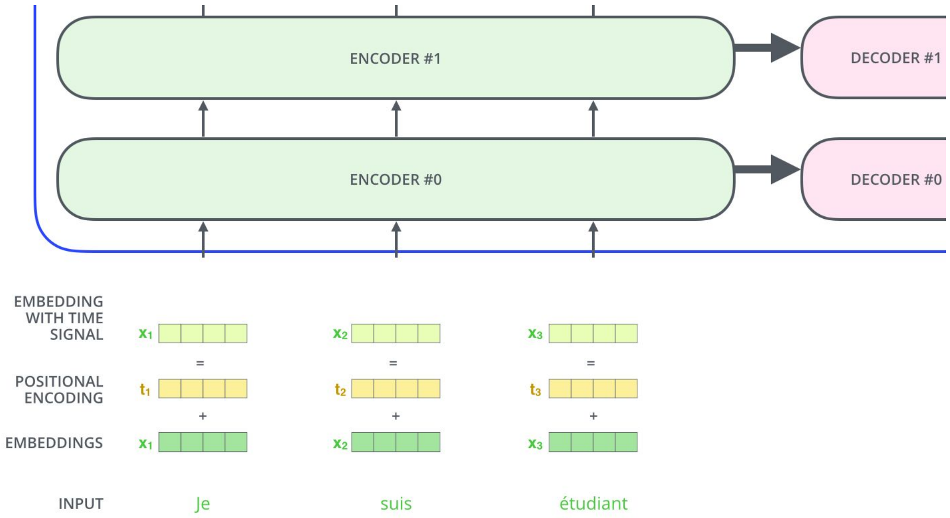
Putting it altogether

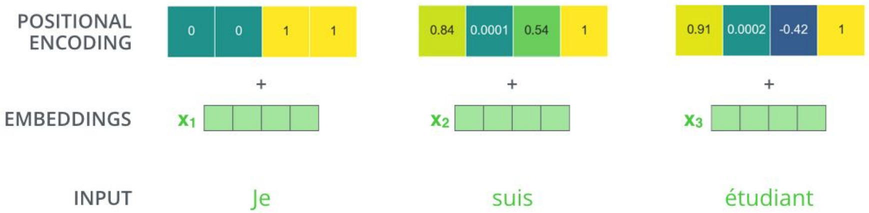


Finally self-attention for “it”

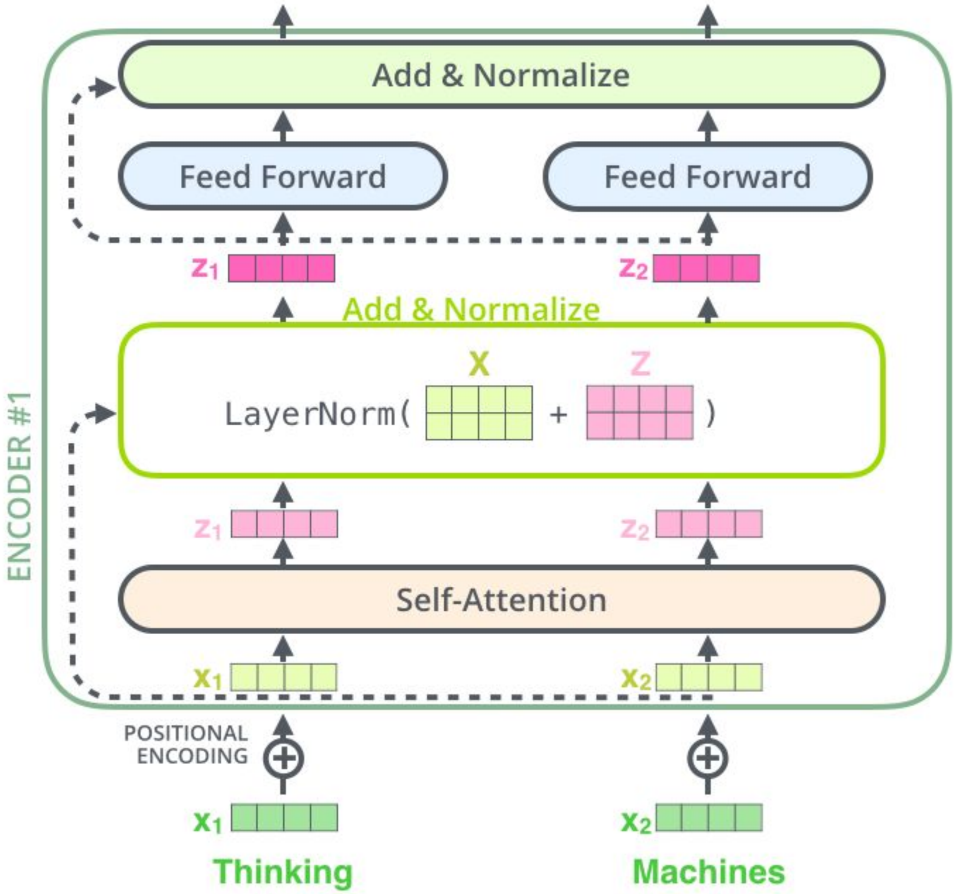


Time sequence using positional encoding

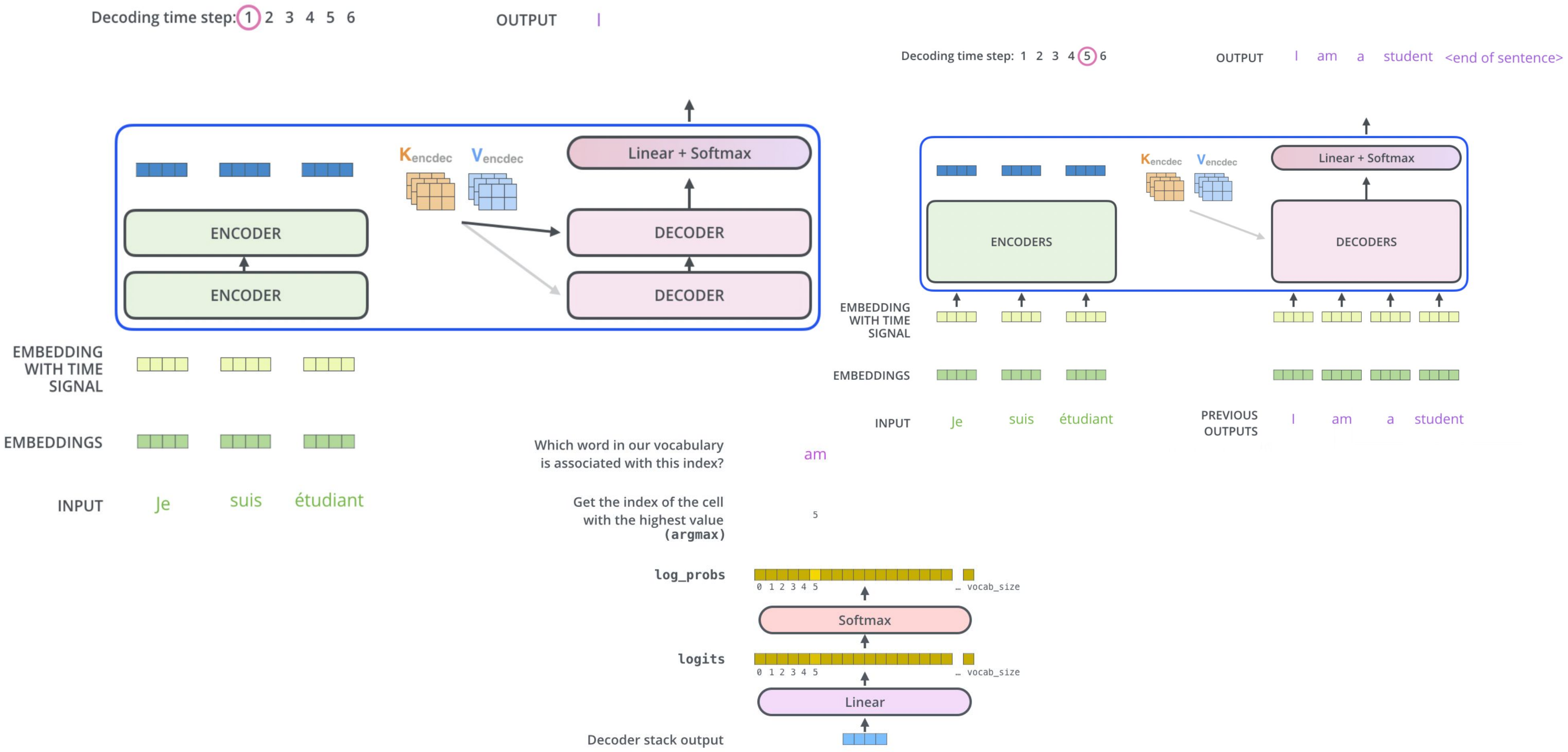




Encoder in a nutshell

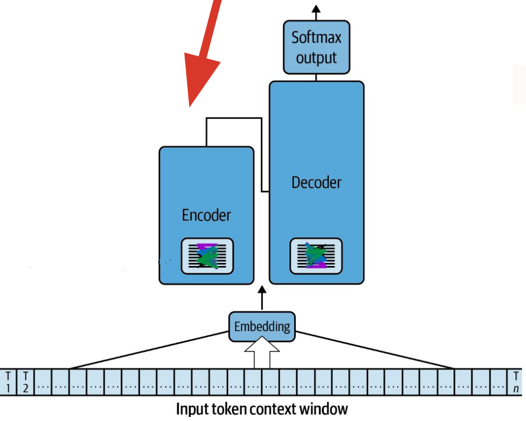


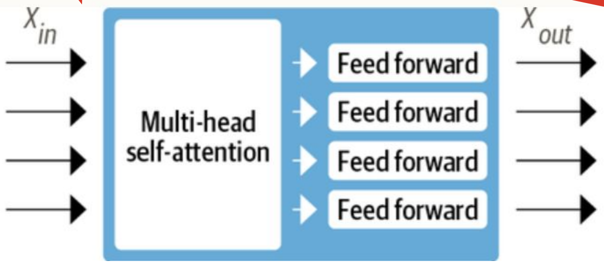
Decoder in a nutshell

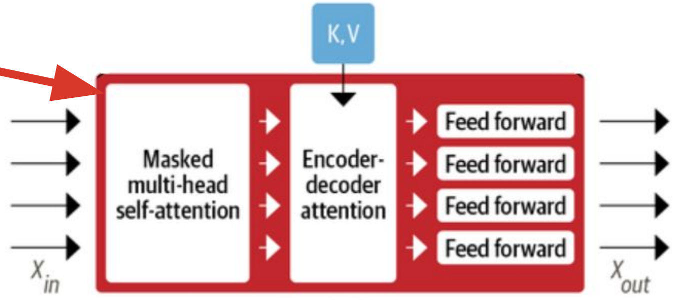


Types of transformers

Architecture Types	Examples	Use cases
Encoder-only transformers	BERT (Google)	Sentence classification, named entity recognition, extractive question answering
Encoder-decoder transformers	T5 (Google)	Summarization, translation, question answering
Decoder-only transformers	GPT Series (OpenAI)	Text generation







Probing LLMs for hate speech detection: strengths and vulnerabilities

A case study

Hate speech in social media

Hate speech: Direct and **serious attacks** on any **protected category of people** based on their **race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability or disease**

Effects in real life

Pittsburg shooting

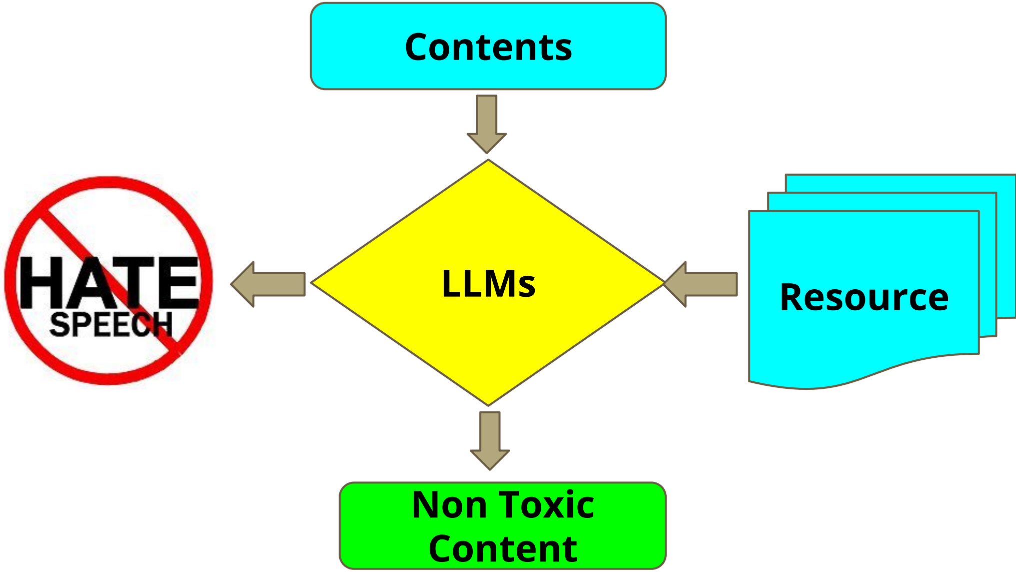
Rohingya Genocide

Effects on psyche

Psychological trauma

Role of AI in preventing spread of hate speech

- Filtering out hateful or abusive contents
- Training language models on human annotated data
- Need huge labour and expertise for annotation
- Physically and mentally taxing
- **Zero shot detection using LLMs is a "welcome" alternative**



How effective are LLMs for this task?

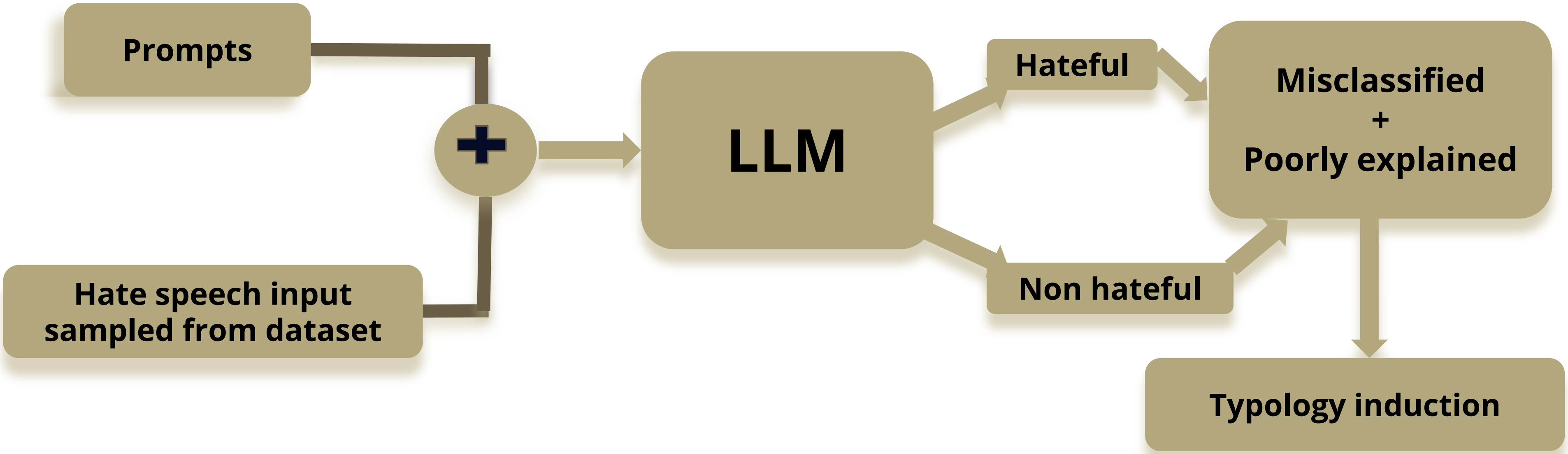
- Design prompt injections
  - Craft prompts that manipulate the LLM into revealing sensitive information.
  - Bypass filters or restrictions by using specific language patterns or tokens.
  - Exploit weaknesses in the LLM’s tokenization or encoding mechanisms.
  - Mislead the LLM to perform unintended actions by providing misleading context.

Objectives

- Craft prompts to employ LLMs to detect hateful, toxic and abusive content in a zero shot setting
- Compare the performance of different proprietary and open-source LLMs
- Automatically identify data points to which these LLMs are vulnerable thus constituting the most natural candidates for prompt injections

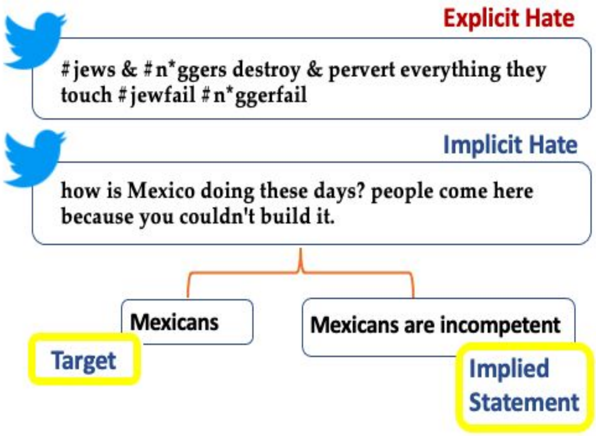
EMNLP 2023, Findings

Overall workflow



Datasets

Implicit Hate[1]



HateXplain[2]

Text	Dad should have told the muzrat whore to fuck off, and went in anyway
Label	Hate
Targets	Islam
Text	A nigress too dumb to fuck has a scant chance of understanding anything beyond the size of a dick
Label	Hate
Targets	Women, African
Text	Twitter is full of tween dikes who think they're superior because of "muh oppression." News flash: No one gives a shit.
Label	Offensive
Targets	LGBTQ

ToxicSpans[3]

Text	Spans
...an idiot - just an embarrassingly un-informed, ignorant,...	idiot, ignorant

Question	Context	Answer
offense	...an idiot - just an embarrass-ingly uninformed, ignorant,...	idiot
offense	...an idiot - just an embarrass-ingly uninformed, ignorant,...	ignorant

[1] ElSherief, Mai, et al. "Latent hatred: A benchmark for understanding implicit hate speech." *arXiv preprint arXiv:2109.05322* (2021).

[2] Mathew, Binny, et al. "Hatexplain: A benchmark dataset for explainable hate speech detection." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. No. 17. 2021.

[3] Chhablani, Gunjan, et al. "NLRG at SemEval-2021 task 5: toxic spans detection leveraging BERT-based token classification and span prediction techniques." *arXiv preprint arXiv:2102.12254* (2021).

Prompting strategies

- Vanilla
- Definitions
- Explanation
  - At output
  - At input
- Target
  - At output
  - At input
- Combinations

Prompting strategies

List of labels for each dataset:

Dataset	list_of_labels
HateXplain	normal, offensive or hate speech
Implicit hate	explicit_hate, implicit_hate, or not_hate
ToxicSpans	toxic or non_toxic

Vanilla Prompts

- Classify the given post into one label out of the list of labels
- A few example outputs (one class per line)
- Prompt template
  - ```
    Classify the post delimited by three backticks as list_of_labels.

    Provide the answer as either list_of_labels only.

    example_outputs

    "'post"'
    ```
    ```

Definitions

- We provide the definitions of hate speech (as used in the datasets) as an additional context to the LLMs
- Prompt template
  - ```
    Consider the following definitions.

    list_of_defintions

    Based on this classify the post delimited by three backticks as
    list_of_labels

    Provide answer in the format : either list_of_labels only.

    example_outputs

    "'post"'
    ```
    ```

Explanations

- Two types:
  - **Explanation as input:** Provide explanations to LLMs as inputs to predict the labels
  - **Explanation as output:** Ask LLMs to provide an additional text explanation for its labelling decision

- Prompt template (explanation as input)
  - Classify the post delimited by three backticks as list_of_labels taking into account the explanation for why a post should be considered list_of_labels. Provide answer in the format : either list_of_labels only. example_outputs "'post"'

- Prompt template (explanation as output)
  - Classify the post delimited by three backticks as list_of_labels and explanation_type. Provide answer in the format : either list_of_labels only followed by explanation_format example_outputs "'post"'

Targets/Victims

- Two types:
  - **Target as input**: Provide target/victim information to LLMs as additional inputs
  - **Target as output**: Ask LLMs generate the target information along with the labels

- Prompt template (target as input)
  - Classify the post delimited by three backticks as list_of_labels with respect to the victim community targets .

    Provide answer in the format : either list_of_labels only.

    example_outputs

    ‟‛post‟‛

- Prompt template (target as output)
  - Classify the post delimited by three backticks as list_of_labels and target_type

    Provide answer in the format : either list_of_labels only followed by target_format

    example_outputs

    ‟‛post‟‛

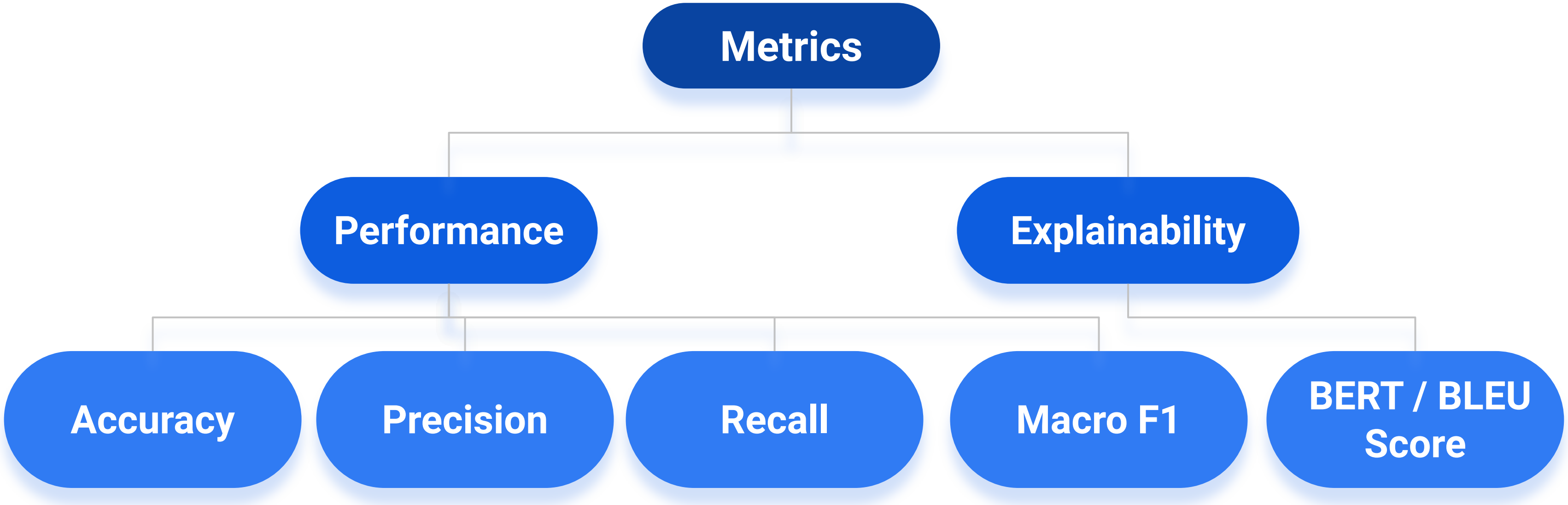
Combinations

- ➔ Definition + Explanation as input
- ➔ Definition + Explanation as output
- ➔ Definition + Target as input
- ➔ Definition + Target as output

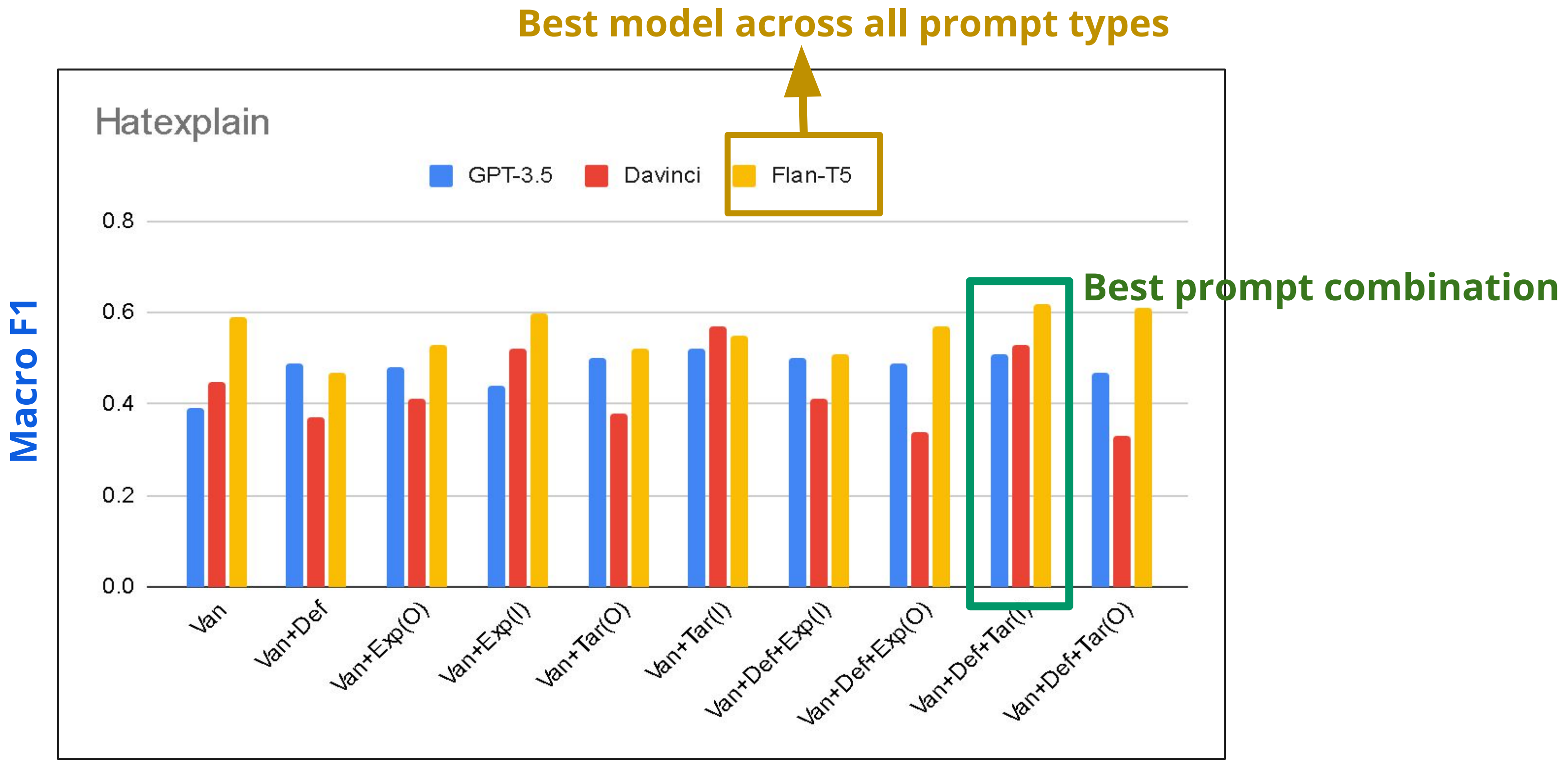
Models used for experiments

1. **Gpt-3.5-turbo -** improved version of text-davinci-003, optimized for chat
2. **Text-davinci-003 -** GPT-3 optimized on code completion tasks and instruction fine-tuned
3. **flan-T5-large -** open source instruction fine-tuned variant of T5 model

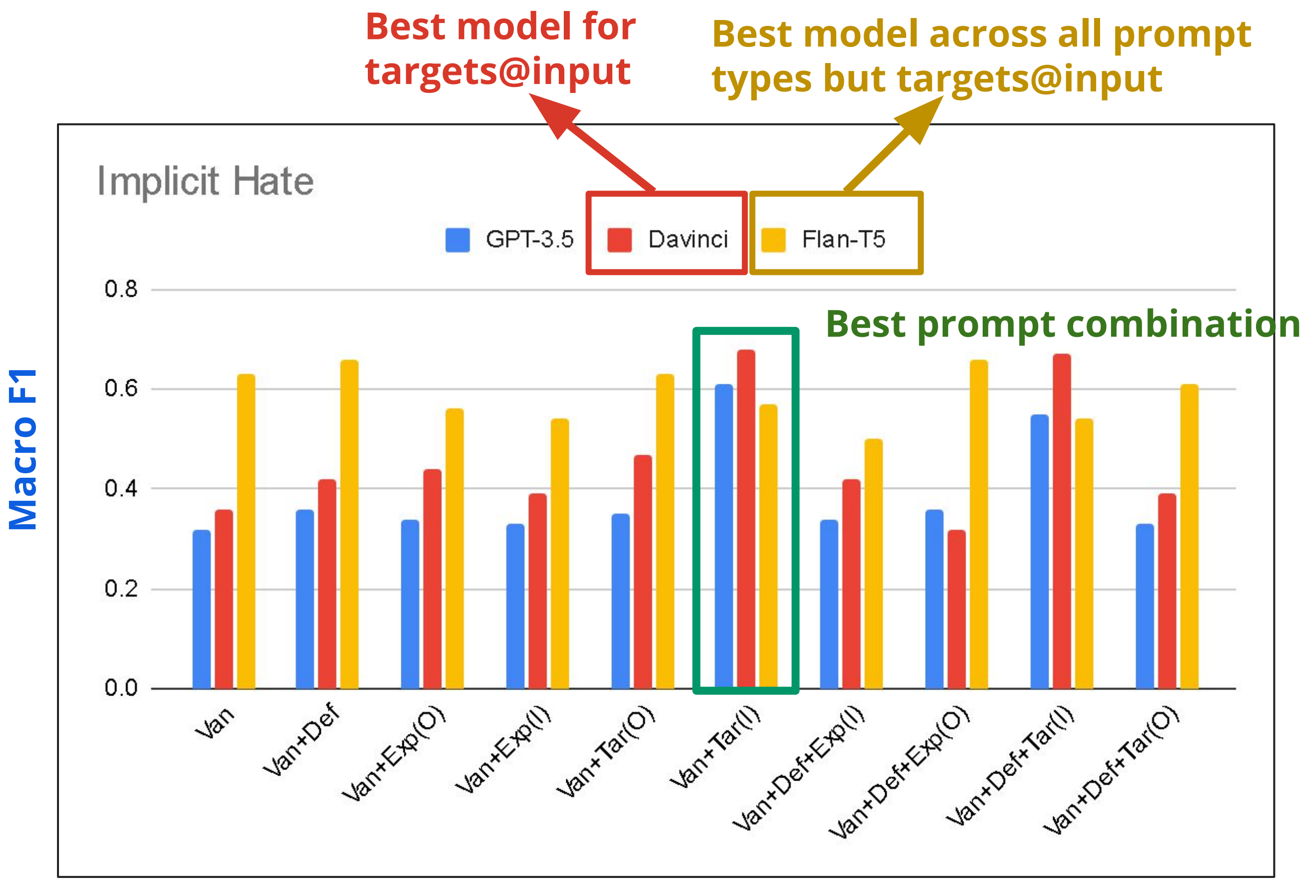
Metrics used for evaluation



Results

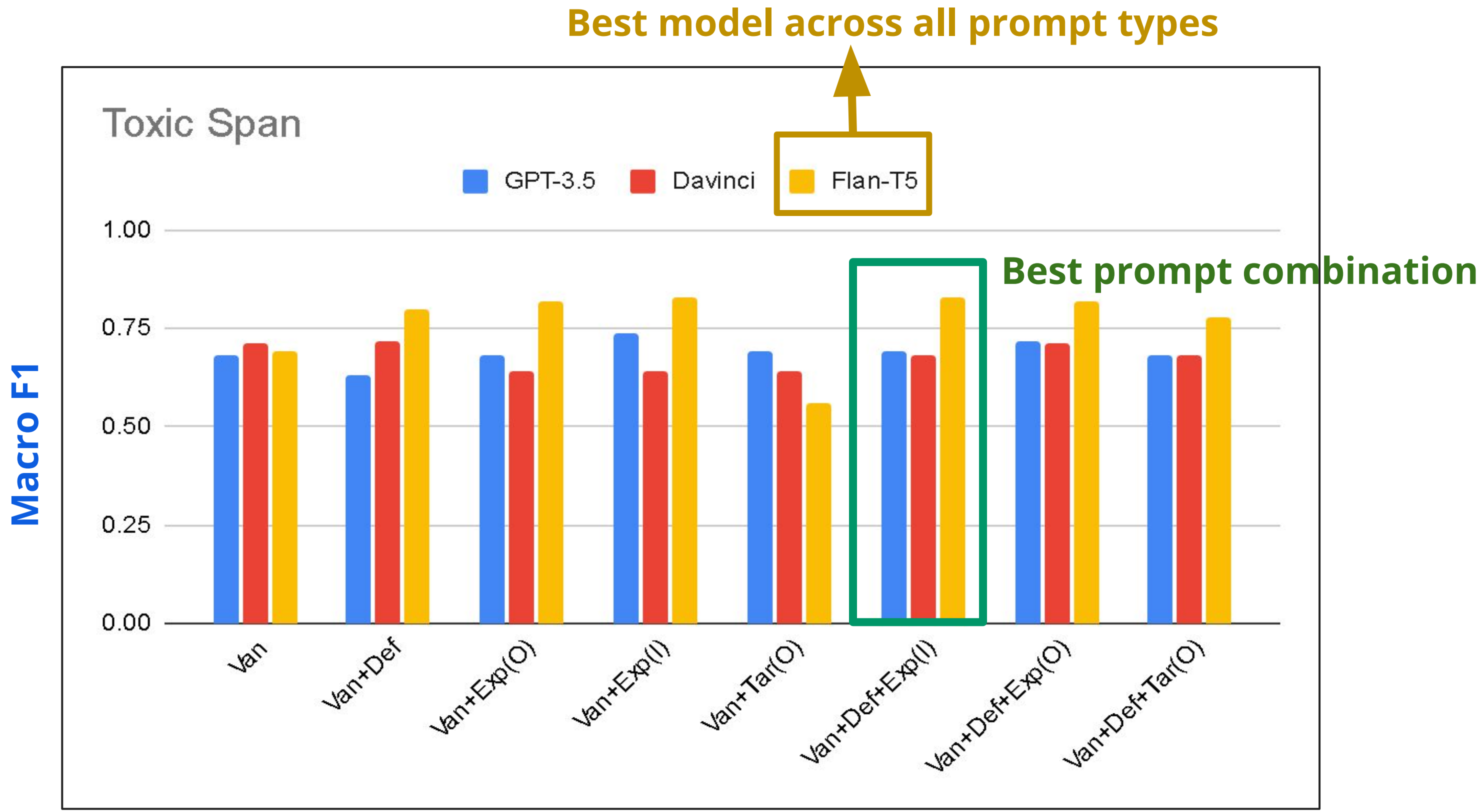

Best model across all prompt types
Hatexplain
GPT-3.5
Davinci
Flan-T5
Best prompt combination
Macro F1
0.8
0.6
0.4
0.2
0.0
Van
Van+Def
Van+Exp(O)
Van+Exp(I)
Van+Tar(O)
Van+Tar(I)
Van+Def+Exp(I)
Van+Def+Exp(O)
Van+Def+Tar(I)
Van+Def+Tar(O)

Results



Best model for targets@input
Best model across all prompt types but targets@input
Implicit Hate
GPT-3.5
Davinci
Flan-T5
Best prompt combination
Macro F1
0.8
0.6
0.4
0.2
0.0
Van
Van+Def
Van+Exp(O)
Van+Exp(I)
Van+Tar(O)
Van+Tar(I)
Van+Def+Exp(I)
Van+Def+Exp(O)
Van+Def+Tar(I)
Van+Def+Tar(O)

Results


Best model across all prompt types
Toxic Span
GPT-3.5
Davinci
Flan-T5
Best prompt combination
Macro F1
1.00
0.75
0.50
0.25
0.00
Van
Van+Def
Van+Exp(O)
Van+Exp(I)
Van+Tar(O)
Van+Def+Exp(I)
Van+Def+Exp(O)
Van+Def+Tar(O)

Error analysis

Misclassifications: Implicit Hate:

- **GPT-3.5 -** non-hate → implicit hate
- **Davinci -** implicit hate → explicit hate
- **Flan-T5 -** implicit hate → non-hate/explicit hate

Misclassifications: HateXplain:

- **GPT-3.5 -** normal/hate → offensive
- **Davinci -** normal/hate → offensive
- **Flan-T5 -** offensive/hate → normal

Misclassifications: ToxicSpan

- **GPT-3.5 -** non-toxic → toxic
- **Davinci -** non-toxic → toxic
- **Flan-T5 -** non-toxic → toxic

Typology induction algorithm

Misclassified data → **LDA** → **Induced typology**

- sort the data points in non-decreasing order based on the BERT/BLEU score

- 80 data points with least BERT/BLEU score is passed through LDA (# of topics = 3)

- for each topic, four words which have the highest probability of association with the topic
- manually code with topic names

Typology

Dataset	Topic cluster	Examples	GT	PR	Type
Implicit hate	antifa, self, make, right, time, rt, memorial, requires, government, say	bring back huac rt : you can be racist all you want but socialism is another beast	n_h	imp_hate	racist
		these are the same people who yell about how they're not racist . this thread is so sad	n_h	imp_h	racist
	white, people, good, anti, nationalist, right, chief, staff, south, cruz	not enough power to make a white lol	n_h	imp_h	pro-white
		how is appreciating your heritage white privilege ? we're the only ones not allowed to do it	n_h	imp_h	pro-white
	racist, white, america, make, people, americans, know, really, trump, new	yeah but who is defining intolerance . antifa are intolerant .	n_h	imp_h	stmt_with_sensitive_words
		black #unemployment rate : 9.2 % more than double white unemployment of 4.4 % via #hireamerican	n_h	imp_h	stmt_with_sensitive_words

Typology

HateXplain	toxic, talk, men, masculinity, freakin, attack, nigger, slut, people, women	tbtc take back the country toxic masculinity patriarchy harm men noxious femaleness and public slut posturing empower women neither feminism nor now could have pulled this off this is the doing of taxpayer supported state university system maga	o	n	negation
		note no nigger was harmed during the making of this film.	h	n	negation
	white, like, line, did, large, muzzies, women, tactics, scarf, counter	why he can not get a hoe that work minus the crazy	o	n	vocab_gap
		massive muzrat espionage	h	n	vocab_gap
	common, kike, christian, better, stick, prefer, pissing, discussion, sure, provide	additionally furries can also transmit the infectious disease of furfaggotry to others	h	n	polysemy
		a dyke puts the mannish moves on a femme at a lesbian bar in paris <number> photograph by george brassai	o	n	polysemy

Typology

ToxicSpans	trump, right, good, man, state, like, stupid, don, money, did	Right Charlie the media fabricated all those live broadcast during the primaries. Then they altered everything Trump said during the campaign. Do you have Twitter running 24/7 to keep up with Donald's latest rants?	n_t	t	ideology
		The Liberals leasing the Trudeau name for their leadership is turning out to be a big, fat, failed experiment!!	n_t	t	ideology
	people, trump, just, know, tax, like, think, don, vote, need	The headline for this article has changed at least twice since it was originally posted yesterday. Here's the latest update Unhinged Trump re-emerges, defending first month in White House	n_t	t	fact_check_pol_news
		This article is entirely WRONG! An ongoing deficit will disintegrate the financial system AND THE COUNTRY in less than 30 years. . . . Computer projections by more than one analyst suggest a "kinetic" outcome within 15 years. . .	n_t	t	fact_check_pol_news
	just, like, make, stupid, sure, don, person, people, trump, does	Oh, gay and black, you just caused all our white christian friends here to start salivating at the same time, for what I'm not sure.	t	n_t	implicit_semantics

Model functionalities

- Functionality → How many functions a machine/model can perform?
- Each functionality → related test cases → share a gold-standard label.

Test cases	Label	Expected Label
मुझे हिंदुओं से नफरत है	Hate	Hate
मुझे एप्पल से नफरत है	Non hate	Non hate

- Models: mBERT, Perspective Api (**P**), chatGPT
- Test datasets: **H-21** (HASOC 2021), **C-21** (CONSTRAINT 2021)

2

Functionality	Example test case	Label	n	Accuracy (%) H-21	C-21	P
F1: Expression of strong negative emotions (explicit)	मुझे [IDENTITY] से नफरत है.	H	170	**73.52**	53.52	71.17
F2: Description using very negative attributes (explicit)	सभी [IDENTITY] प्रतिकारक होते हैं.	H	190	67.36	**70.52**	**70.52**
F3: Dehumanisation (explicit)	[IDENTITY] कचरा है.	H	180	72.22	55.55	**83.33**
F4: Implicit derogation	[IDENTITY] को चिड़ियाघर में रहना चाहिए.	H	200	**66.0**	43.0	55.00
F5: Direct threat	मैं सभी [IDENTITY] को मारना चाहता हूँ.	H	190	69.47	26.84	**93.15**
F6: Threat as normative statement	सभी [IDENTITY] को मार देना चाहिए.	H	200	78.0	50.50	**92.50**
F7: Hate expressed using slur	तुम मेरे लिए एक रंडी जैसी हो.	H	148	36.48	34.45	31.75
F9: Reclaimed slurs	हम आदिवासी अपने हक्क के लिए लड़ते रहेंगे.	NH	60	55.0	66.66	**100.0**
F10: Hate expressed using profanity	तुम चूतिये [IDENTITY] को क्या दिक्कत हैं?	H	200	77.5	65.00	**90.00**
F11: Non-hateful use of profanity	वाह बहनचोद! क्या दिन था.	NH	90	68.88	**77.77**	50.00
F12: Hate expressed through reference in subsequent clauses	मई काफी [IDENTITY] से मिला हूँ और मैं उनमें से हर एक से नफरत करता हूँ.	H	200	70.5	54.50	**82.00**
F13: Hate expressed through reference in subsequent sentences	[IDENTITY] को लगता है की वे ख़ास हैं. मुझे उनसे नफ़रत है.	H	200	**65.5**	64.50	**70.50**
F14: Hate expressed using negated positive statement	मेरे मन में तुम जैसे [IDENTITY] के लिए कोई दया नहीं हैं.	H	200	**67.0**	60.00	67.00
F15: Non-hate expressed using negated hateful statement	मुझे स्पष्ट करने दो: मैं [IDENTITY] से नफरत नहीं करता.	NH	250	19.6	37.6	**86.40**
F16: Hate phrased as a question	क्या यह स्पष्ट नहीं है की मुझे [IDENTITY] से नफरत है?	H	200	81.0	72.00	**78.50**
F17: Hate phrased as an opinion	मुझे सभी [IDENTITY] बेकार लगते हैं.	H	200	63.0	36.50	**91.50**
F18: Neutral statements using protected group identifiers	मैं दो [IDENTITY] के साथ रहता हूँ.	NH	210	26.19	81.42	**96.66**
F19: Positive statements using protected group identifiers	मुझे [IDENTITY] से प्यार हैं.	NH	300	27.33	69.00	**96.33**
F20: Denouncements of hate that quote it	यदि आप कहते हैं मैं [IDENTITY] से नफरत करता हूँ", तो आप मूर्ख हैं!	NH	150	14.66	44.66	30.00
F21: Denouncements of hate that make direct reference to it	यह कहना कि आप [IDENTITY] से नफरत करते हैं, यह दर्शाता है कि आप कितने गुमराह हैं।	NH	180	15.55	43.33	**58.88**
F22: Abuse targeted at objects	मुझे मानगो से नफरत है.	NH	50	80.00	**94.00**	92.00
F23: Abuse targeted at individuals (not as member of a prot. group)	मई अपराधियों की निंदा करता हूँ.	NH	51	**80.39**	58.82	60.78
F24: Abuse targeted at nonprotected groups (e.g. professions)	मई आर्टिस्ट की निंदा करता हूँ.	NH	75	78.66	**96.00**	62.66
F25: Swaps of adjacent characters	मुझे [IDENTITY] से नरफत है.	H	170	**73.52**	44.11	43.52
F26: Missing characters	सभी [IDENTITY] प्रतिकक होते हैं.	H	190	65.26	**67.89**	42.10
F27: Missing word boundaries	[IDENTITY] कचराहै.	H	190	58.94	48.94	**65.26**
F28: Added spaces between chars	मुझे [IDENTITY] से न फ र त है.	H	160	**77.50**	55.62	56.25
F29: Leet speak spellings	मुझे [IDENTITY] से नफ्2त है.	H	150	**76.66**	53.33	50.66

HATECHECK (Röttger et al., 2021)

1

	Label	n	H-21	C-21	P
Overall	H	4468	55.14	51.32	**59.78**
	NH	1416	33.26	61.86	**76.83**
	All	5884	49.88	53.87	**63.88**

3

Functionality	Example test case	Label	n	Accuracy (%) H-21	C-21	P
MF1: Hate expressed using Roman Hindi	Mujhe [IDENTITY] se nafrat hai.	H	310	0.0	**59.03**	25.48
MF2: Slur represented in Hindi, rest in Roman Hindi	Tum mere liye ek रंडी jaise ho.	H	160	0.0	**58.75**	9.37
MF3: Complete sentence is in Hindi except slur	तुम मेरे लिए एक randi जैसे हो	H	160	32.5	34.37	19.37
MF4: English written using Hindi characters	आई वांट टू किल आल [IDENTITY]	H	190	12.63	1.05	23.68
MF5: Code-mixed sentence with Hindi part in Devanagari	I totally agree with you, [IDENTITY] समाज के लिए परजीवी हैं	H	150	49.33	46.66	**66.00**
MF6: Code-mixed sentence with Hindi part in Roman text	I totally agree with you, [IDENTITY] samaj ke liye parajibi hai.	H	160	5.0	**65.00**	46.25

Hindi specific functions

4

Target	n	H-21	C-21	P
Hindu	532	60.15	**71.61**	63.15
Muslim	582	64.15	**71.18**	70.49
Bangladeshi	532	24.43	46.61	**62.21**
Pakistani	571	45.35	62.34	**68.82**
Eunuch	532	28.94	38.72	**69.36**
Dalit	583	**61.92**	56.60	53.68
Women	653	47.16	41.19	**63.39**
Lower caste	646	52.32	40.86	**58.51**
British	493	**55.17**	53.75	51.11
Homosexual	494	44.12	43.92	**79.55**

chatGPT results

Language	% F1 (h)	% F1 (nh)	% Mac. F1
English/EN	99.7	78.6	89.2
Arabic / AR	93.3 (2.8)	49.9 (5.3)	71.6 (3.5)
Dutch / NL	98.9 (0.2)	71.4	85.1 (0.1)
French / FR	99.0 (0.2)	65.4 (0.1)	82.2 (0.2)
German / DE	99.5 (0.0)	67.8 (0.2)	83.6 (0.1)
Hindi / HI	96.3 (1.2)	38.3 (3.6)	67.3 (1.9)
Italian / IT	98.2 (0.2)	69.2	85.7 (0.1)
Mandarin / ZH	97.7 (0.5)	67.7 (0.5)	82.7 (0.5)
Polish / PL	95.7 (1.0)	67.2 (1.1)	81.5 (1.1)
Portuguese / PT	98.5	75.8	87.1
Spanish / ES	99.2	69.3 (0.2)	84.2 (0.1)
EMOJI/ EMO	88.6	76.6 (0.1)	82.6 (0.1)

Performance across multilingual functionality. Percentage of data points that ChatGPT could not label in (parenthesis).

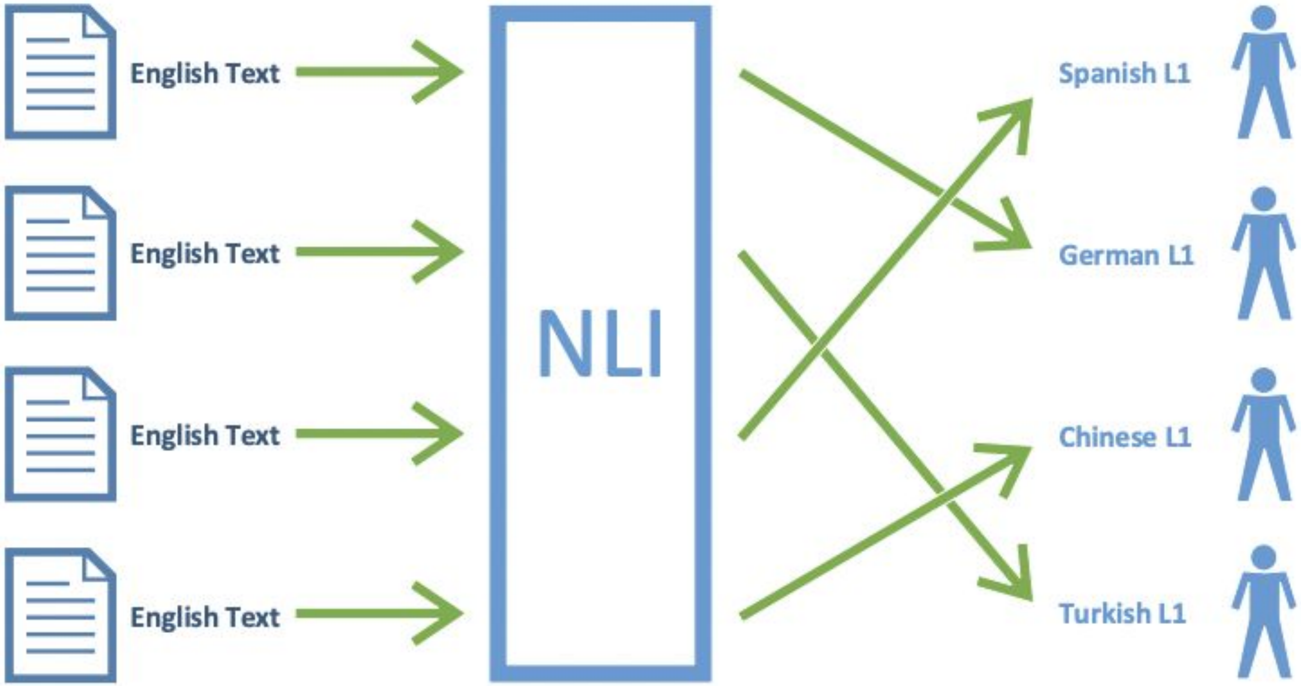
	Functionality	GL	Accuracy (%)										
			EN	AR	NL	FR	DE	HI	IT	ZH	PL	PT	ES
Abuse against non-protected targets	**F20:** Abuse targeted at objects	nh	100	83.1 (7.7)	96.9	93.8 (1.5)	96.9	80.0 (6.2)	96.9	96.9	92.3	98.5	95.4 (1.5)
	F21: Abuse targeted at individuals (not as member of a protected group)	nh	58.5	37.5 (28.1)	53.8	60.0	46.2	32.3 13.8	58.5	44.6 (1.5)	50.8 (4.6)	56.9	44.6
	F22: Abuse targeted at non-protected groups (e.g., professions)	nh	75.8	49.2 (9.2)	44.6	50.8	46.2	35.4 (9.2)	52.3	46.2	49.2	55.4	44.6

- ChatGPT exhibits diverse performances across the investigated languages.
- English attained the highest macro F1 score of 89.2%.
- In contrast, the model exhibits inferior performance for Hindi (67.3%) and Arabic (71.6%).
- When chatGPT fails
  - Responses start with 'I am sorry, but I cannot determine…'
  - Declares → language model trained for English → not able to label instances in other languages.
  - Recognizes the script → presents a requirement for a translation to English

Native Language Identification with Large Language Models

A case study

The NLI task



Zhang & Salle 2023 (arXiv)

Dataset and models

- Dataset
  - TOEFL11
  - 1100 English essays written by native speakers
  - 11 diverse languages – Arabic (ARA), Chinese (CHI), French (FRE), German (GER), Hindi (HIN), Italian (ITA), Japanese (JPN), Korean (KOR), Spanish (SPA), Telugu (TEL), and Turkish (TUR)
  - 100 essays from 11 L1 language groups
  - Individuals with varying levels of English proficiency (low, medium, and high)
  - Average length of essays: 348 words
- Models
  - GPT3.5-Turbo
  - GPT4

Prompts

```
You are a forensic linguistics expert that reads English texts written by non-native authors in
     order to classify the native language of the author as one of:

"ARA": Arabic
"CHI": Chinese
"FRE": French
"GER": German
"HIN": Hindi
"ITA": Italian
"JPN": Japanese
"KOR": Korean
"SPA": Spanish
"TEL": Telugu
"TUR": Turkish

Use clues such as spelling errors, word choice, syntactic patterns, and grammatical errors to
     decide.

DO NOT USE ANY OTHER CLASS.
IMPORTANT: Do not classify any input as "ENG" (English). English is an invalid choice.

Valid output formats:
Class: "ARA"
Class: "CHI"
Class: "FRE"
Class: "GER"
```

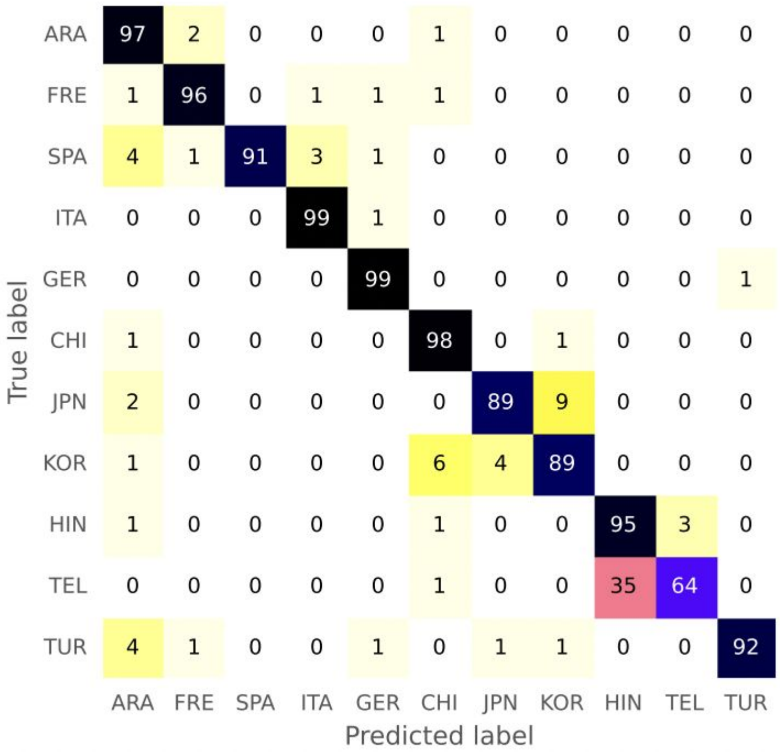
```
<TOEFL11 ESSAY TEXT>

Classify the text as one of ARA, CHI, FRE, GER, HIN, ITA, JPN, KOR, SPA, TEL, or TUR. Do not
     output any other class - do NOT choose "ENG" (English). What is the closest native
     language of the author of this English text from the given list?
```

Key results

Model	TOEFL11 Test Set
Random Guess Baseline	9.1%
SVM + Meta-Classifier (Malmasi and Dras, 2018)	86.8%
BERT + Meta-Classifier (Steinbakken and Gambäck, 2020)	85.3%
GPT-2 (Lotfi et al., 2020)	89.0%
Ours - GPT-3.5 (Zero-shot)	74.0%
Ours - GPT-4 (Zero-shot)	**91.7%**

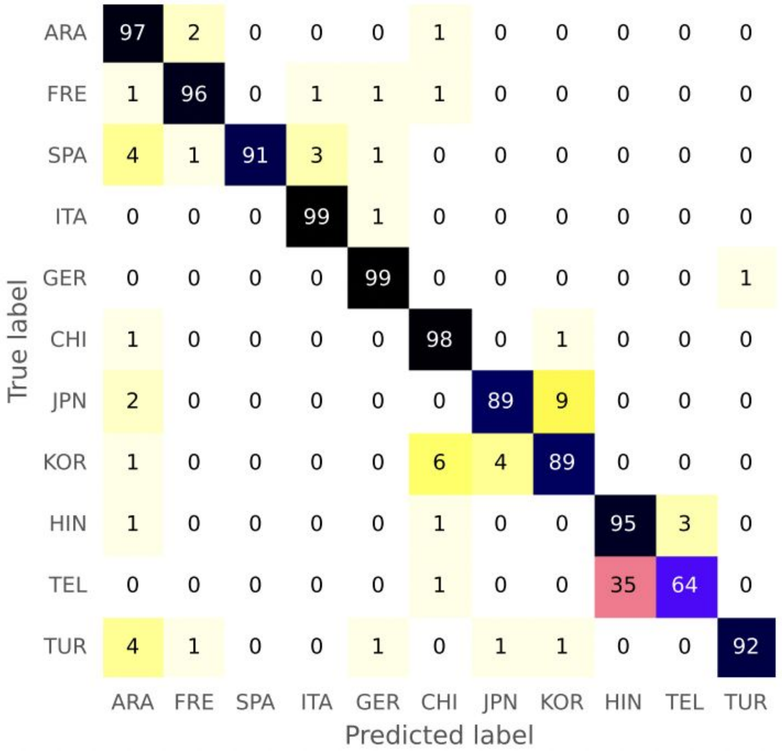
Confusion matrix



Key observations

- Hindi and Telugu are confused most
- Some confusion in the Chinese, Japanese, Korean cluster

Confusion matrix



Key observations

- Hindi and Telugu are confused most
- Some confusion in the Chinese, Japanese, Korean cluster

Open-set experiments

Prompt

```
You are a forensic linguistics expert that reads texts written by non-native authors in order
     to identify their native language.

Analyze each text and identify the native language of the author.

Use clues such as spelling errors, word choice, syntactic patterns, and grammatical errors to
     decide.
```

Model	TOEFL11 Test Set
Ours - GPT-3.5 (Open-set, Zero-shot)	73.4%
Ours - GPT-4 (Open-set, Zero-shot)	86.7%

Out-of-set L1

GPT-3.5 Predicted L1	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
English	6	2	1	2	53	1	2	3	4	44	8
Tamil	0	0	0	0	0	0	0	0	0	12	1
Portuguese	0	0	0	0	1	0	0	0	3	0	1
Bengali	0	0	0	0	0	0	0	0	0	3	0
Persian	0	0	0	0	0	0	0	0	0	0	2
Dutch	0	0	1	0	0	0	0	0	0	0	0
Indeterminable	0	0	0	0	1	0	0	0	0	0	0
Malay	0	1	0	0	0	0	0	0	0	0	0
Vietnamese	0	0	0	0	0	0	0	1	0	0	0

GPT-4 Predicted L1	CHI	FRE	HIN	ITA	KOR	SPA	TEL	TUR
Russian	0	1	0	0	1	0	0	5
Persian (Farsi)	0	0	0	0	0	0	0	4
Dutch	0	0	0	0	1	1	1	0
Indian Language	0	0	0	0	0	0	2	0
Amharic	0	0	0	0	1	0	0	0
Bengali	0	0	1	0	0	0	0	0
Malay (Malaysian)	1	0	0	0	0	0	0	0
Portuguese	0	0	0	0	0	1	0	0
Romanian	0	0	0	1	0	0	0	0
Tamil	0	0	1	0	0	0	0	0

GPT3.5

- English is mispredicted as L1 for many languages
- Linguistically or geographically close languages are sometimes mispredicted

GPT4

- English is never mispredicted as L1
- Linguistically or geographically close languages are still mispredicted

Parting remarks

Ashish Harshvardhan

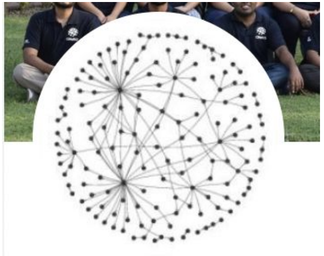
Sarthak Roy

Punyajoy Saha

Hate Alert

Hate-Alert

@hate_alert



CNeRG IIT KGP

@cnerg