ACL 2023

Pre-trained Language Models Can be Fully Zero-Shot Learners

Xuandong Zhao[1], Siqi Ouyang[1], Zhiguo Yu[2], Ming Wu[2], Lei Li[1]

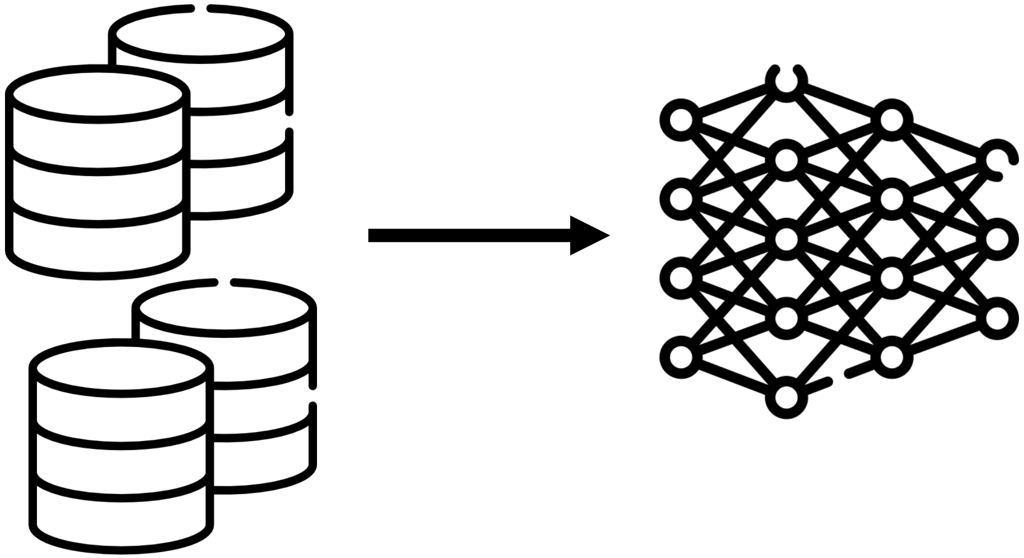
[1]UC Santa Barbara [2]Microsoft

Large language model is a game-changer

Language understanding with LLM

Fine-tuning

- Requires large amounts of labeled data
- Computationally expensive



Language understanding with LLM

In-context learning

- Sensitive to the choice of few-shot demonstrations
- Scales poorly to large test sets (consumes lots of tokens)

k Demonstration Examples

E-mail scam targets police chief Wiltshire Police warns about "phishing" after its fraud …	Topic: Science
Storage, servers bruise HP earnings update Earnings per share rise compared with a year …	Topic: Science
Dutch Retailer Beats Apple to Local Download Market AMSTERDAM (Reuters) - Free …	Topic: Science
Super ant colony hits Australia A giant 100km colony of ants which has been discovered in …	Topic: Science
IBM to hire even more new workers By the end of the year, the computing giant plan …	Topic: Science
Sister of man who died in Vancouver police custody slams chief (Canadian Press) …	Topic: Politics
……	…..
Giddy Phelps Touches Gold for First Time Michael Phelps won the gold medal in …	Topic: Sports

Language understanding with LLM

Zero-shot learning

- Requires human effort to select class description tokens

In this task, you are given a sentence. Your job is to classify the following sentence into one of the four different categories. The categories are: “politics”, “sports”, “business”, and “technology”.

The politics category is related to politics, government, and law.
The sports category is related to sports, competition, and athletics.
The business category is related to business, portfolio, economics, and money.
The technology category is related to technology, software, system, and science.

Category Descriptions

Input: British athletics appoint psychologist for 2008 Olympics British athletics chiefs have appointed sports ...

Output:

sports

Language understanding with LLM

How to do fully zero-shot learning setting?

- Only label names are available

Our approach: NPPrompt

Nonparametric prompting for pre-trained language model

- generate predictions for semantic labels with test text alone

NPPrompt: Short prompt

Predict

next token (GPT/T5)
mask token (BERT/RoBERTa)

The Warriors won the NBA championship 2022. | This | topic | is | about | [MASK]

Original input | Template

NPPrompt: Find label-related words

Category

“SPORTS”

Top-k Closest Words to “SPORTS”

sports, Sports, sport, sporting, athletic, athletics, SPORTS, football, soccer, basketball, tennis …, NBA, …

Initial word embedding of LM

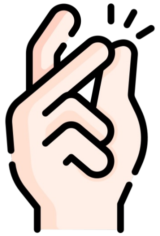
$$S(\mathbf{emb}(v_i), \mathbf{emb}(y_j)) = \frac{\mathbf{emb}(v_i)}{\| \mathbf{emb}(v_i) \|} \cdot \frac{\mathbf{emb}(y_j)}{\| \mathbf{emb}(y_j) \|}$$

$$\mathcal{M}(\text{SPORTS}) = \underset{v_i \in \mathcal{V}}{\text{Top} - k}\{S(\mathbf{emb}(v_i), \mathbf{emb}(\text{SPORTS}))\}$$

NPPrompt: Find label-related words

Does not require **human effort**
Does not require **external knowledge**
Does not require **raw text**

Execute only once

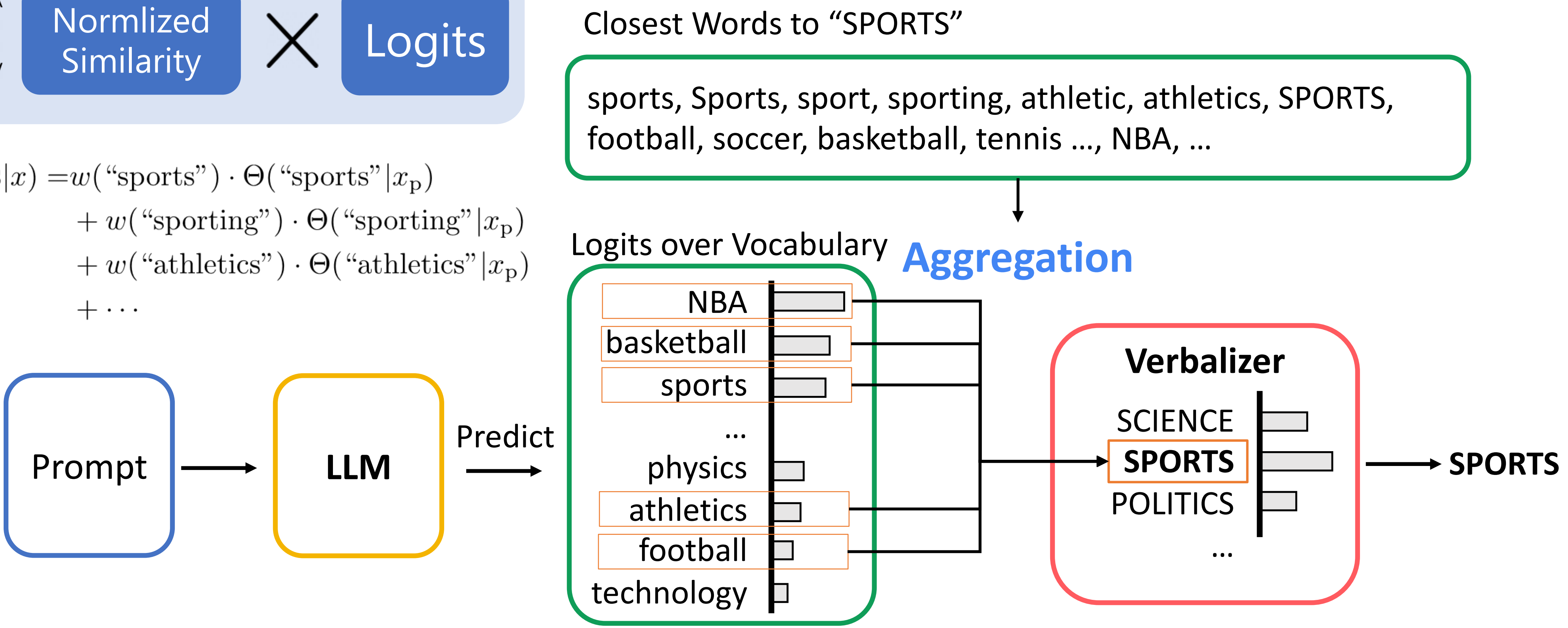


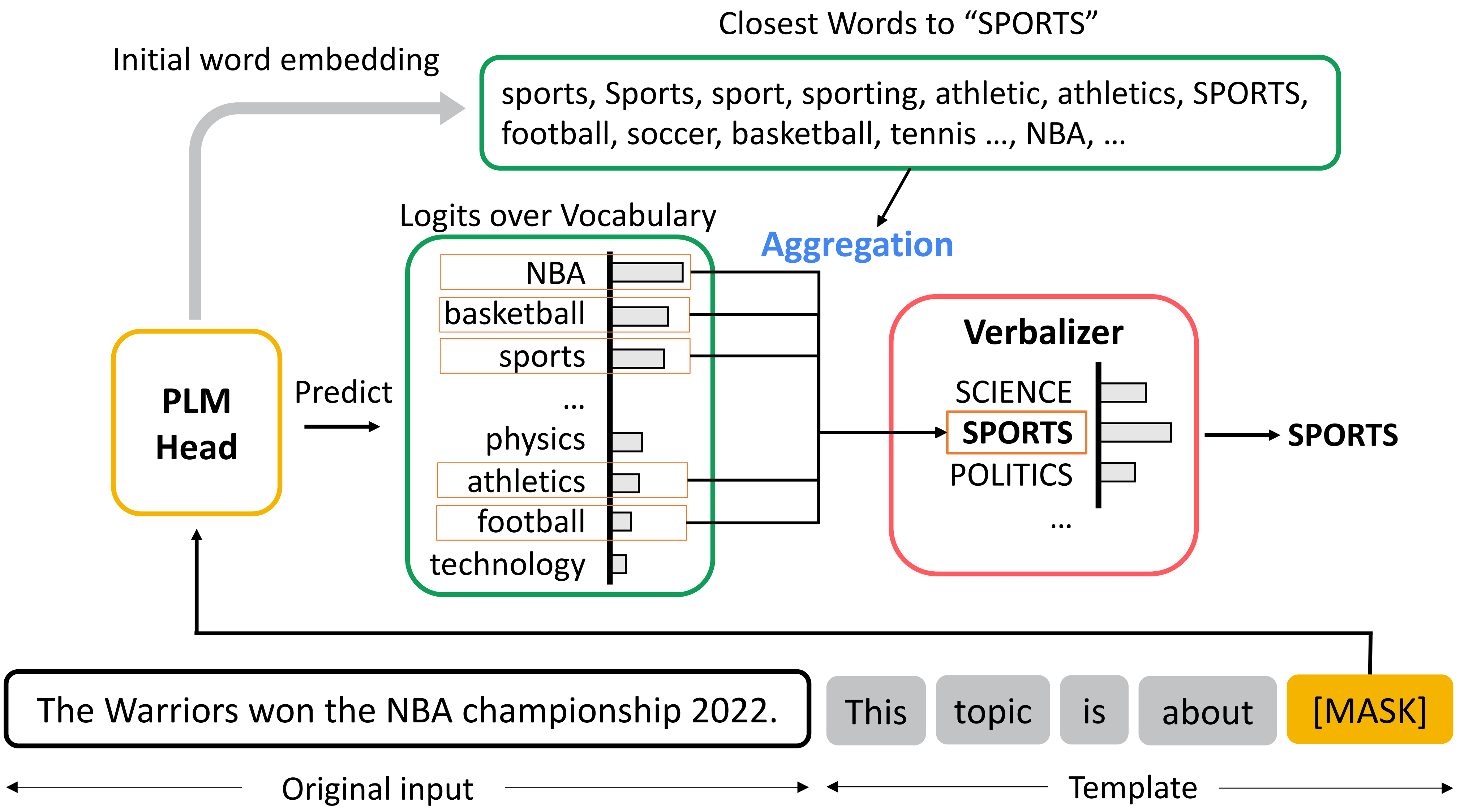
Word	Similarity
“ sports”	1.00
“ Sports”	0.77
“ sporting”	0.68
“ athletics”	0.65
“ athletic”	0.61

NPPrompt: Nonparametric aggregation

$\sum$ Normlized Similarity $\times$ Logits

$$Q(\textsc{Sports}|x) = w(\text{“sports”}) \cdot \Theta(\text{“sports”}|x_{\mathrm{p}}) + w(\text{“sporting”}) \cdot \Theta(\text{“sporting”}|x_{\mathrm{p}}) + w(\text{“athletics”}) \cdot \Theta(\text{“athletics”}|x_{\mathrm{p}}) + \cdots$$




Closest Words to "SPORTS"
Initial word embedding
sports, Sports, sport, sporting, athletic, athletics, SPORTS,
football, soccer, basketball, tennis …, NBA, …
Logits over Vocabulary
Aggregation
NBA
basketball
sports
…
physics
athletics
football
technology
PLM
Head
Predict
Verbalizer
SCIENCE
SPORTS
POLITICS
…
SPORTS
The Warriors won the NBA championship 2022.
This
topic
is
about
[MASK]
Original input
Template

Experiment

Dataset	Classification Type	# Classes
AG News	News Topic	4
DBPedia	Wikipedia Topic	14
IMDB	Movie Review Sentiment	2
Amazon	Product Review Sentiment	2

GLUE benchmark

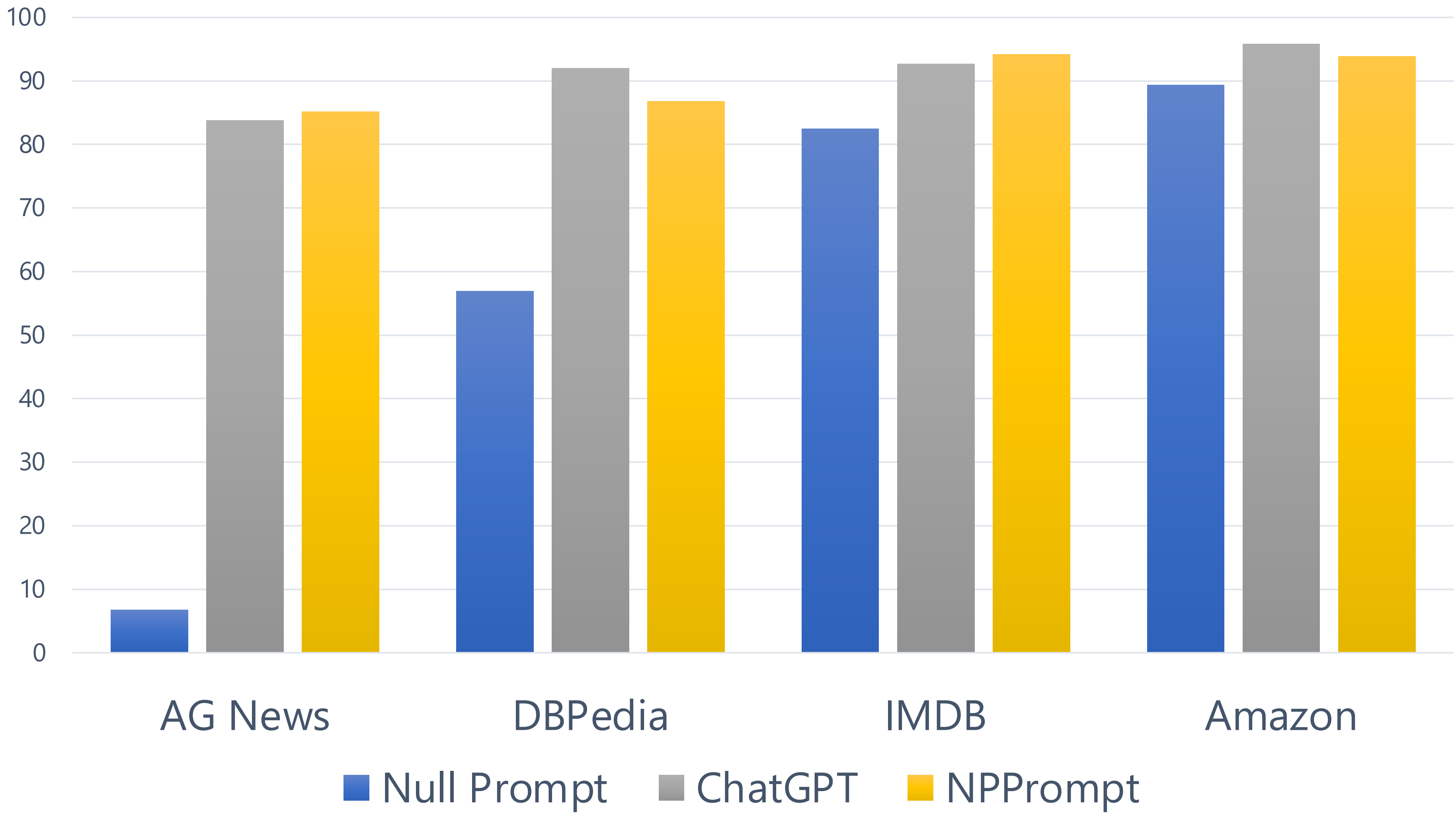
- MNLI, MNLI-mm, SST-2, QNLI, RTE, MRPC, QQP, CoLA

Model: RoBERTa-large

Experiment results

Method	Human/KB	Unlabeled	AG News	DBPedia	IMDB	Amazon	Avg.
ManualVerb	✔	✗	$79.6_{0.6}$	$71.7_{1.1}$	$92.0_{0.7}$	$87.3_{0.4}$	82.7
Semantic Retrieval	✔	✗	$73.1_{1.2}$	$78.6_{0.8}$	$64.8_{1.3}$	$59.4_{0.7}$	69.0
NSP-BERT	✔	✗	$77.4_{0.6}$	$64.7_{5.3}$	$72.8_{1.1}$	$72.7_{3.9}$	71.9
GPT-3 w. descriptions	✔	✗	83.4	82.5	88.8	89.4	86.0
ChatGPT w. descriptions	✔	✗	83.8	92.0	92.7	**95.8**	91.1
SimPTC	✔	✗	$\mathbf{86.9}_{0.3}$	$\mathbf{93.2}_{1.0}$	$91.0_{0.0}$	$93.9_{0.0}$	**91.3**
LOTClass w/o. self train	✗	✔	82.2	86.0	80.2	85.3	83.4
LOTClass	✗	✔	86.4	91.1	86.5	91.6	88.9
KPT	✔	✔	86.7	87.4	**94.0**	94.6	90.7
Null Prompt	✗	✗	$67.9_{2.0}$	$56.8_{3.9}$	$82.5_{1.5}$	$89.4_{1.0}$	74.2
Multi-Null Prompt	✗	✗	$68.2_{1.8}$	$67.6_{1.8}$	$86.6_{0.6}$	$86.2_{2.7}$	77.2
NPPrompt	✗	✗	$\mathbf{85.2}_{0.5}$	$\mathbf{86.8}_{0.1}$	$\mathbf{94.2}_{0.2}$	$\mathbf{93.9}_{0.0}$	**90.0**

Topic/sentiment classification



GLUE benchmark

	MNLI (acc)	MNLI-mm (acc)	SST-2 (acc)	QNLI (acc)	RTE (acc)	MRPC (F1)	QQP (F1)	CoLA (Matt.)	Avg.
With human designed prompts / few-shot data									
Manual Label	50.8	51.7	83.6	50.8	51.3	61.9	49.7	2.0	50.2
In-context learning	$\mathbf{52.0}_{0.7}$	$\mathbf{53.4}_{0.6}$	$84.8_{1.3}$	$53.8_{0.4}$	$60.4_{1.4}$	$45.7_{6.0}$	$36.1_{5.2}$	$-1.5_{2.4}$	48.1
Auto-L	$41.6_{5.4}$	$42.3_{6.2}$	$84.3_{3.3}$	$57.9_{3.9}$	$\mathbf{61.9}_{7.5}$	$\mathbf{67.7}_{7.9}$	$55.5_{5.0}$	$1.2_{4.8}$	51.6
AMuLaP	$50.8_{2.1}$	$52.3_{1.8}$	$\mathbf{86.9}_{1.6}$	$53.1_{2.8}$	$58.9_{7.9}$	$56.3_{5.0}$	$60.2_{2.7}$	$2.3_{1.4}$	52.6
Few-shot fine-tuning	$45.8_{6.4}$	$47.8_{6.8}$	$81.4_{3.8}$	$\mathbf{60.2}_{6.5}$	$54.4_{3.9}$	$76.6_{2.5}$	$\mathbf{60.7}_{4.3}$	$\mathbf{33.9}_{14.3}$	**57.6**
Fully zero-shot									
Majority	32.7	33.0	50.9	49.5	52.7	**81.2**	0.0	0.0	37.5
Null Prompt	$33.1_{0.4}$	$33.8_{0.5}$	$79.1_{4.0}$	$50.7_{0.1}$	$47.2_{0.6}$	$12.9_{7.0}$	$1.3_{1.0}$	$-1.1_{2.0}$	32.1
Multi-Null Prompt	$38.0_{3.5}$	$38.5_{4.1}$	$70.2_{7.7}$	$52.2_{1.7}$	$53.0_{2.2}$	$19.9_{8.7}$	$25.5_{13.4}$	$\mathbf{6.2}_{2.0}$	37.9
NPPrompt	$\mathbf{45.7}_{0.6}$	$\mathbf{45.9}_{0.5}$	$\mathbf{86.3}_{1.2}$	$\mathbf{57.6}_{0.7}$	$\mathbf{55.0}_{3.4}$	$79.8_{1.6}$	$\mathbf{52.4}_{0.4}$	$4.9_{4.1}$	**53.5**

NPPrompt works with different PLMs

Model	AG News	DBPedia	IMDB	Amazon	Average
T5-base	76.8	78.3	68.5	65.3	72.2
GPT2-base	**81.1**	78.1	83.7	**85.6**	82.1
BERT-base	79.4	77.8	57.7	53.5	67.1
RoBERTa-base	75.3	**82.8**	**88.7**	83.9	**82.7**

Summary

- **NPPrompt**, a novel method for **fully zero-shot** learning with pre-trained language models (PLMs).
- NPPrompt utilizes PLMs' **initial word embeddings** to identify related words for category names, without manual design or unlabeled data.
- Empirical results show that NPPrompt significantly outperforms previous fully zero-shot methods.

Takeaway

NPPrompt can be easily plugged into any SOTA LLM

- Employ k-Nearest-Neighbor in LLM's token embedding space
- Nonparametric aggregation
- **Efficient natural language understanding**
- Dynamic zero-shot problems

Thanks for your listening!

SCAN ME

Homepage: https://xuandongzhao.github.io/
Email: xuandongzhao@ucsb.edu

Different neighborhood numbers