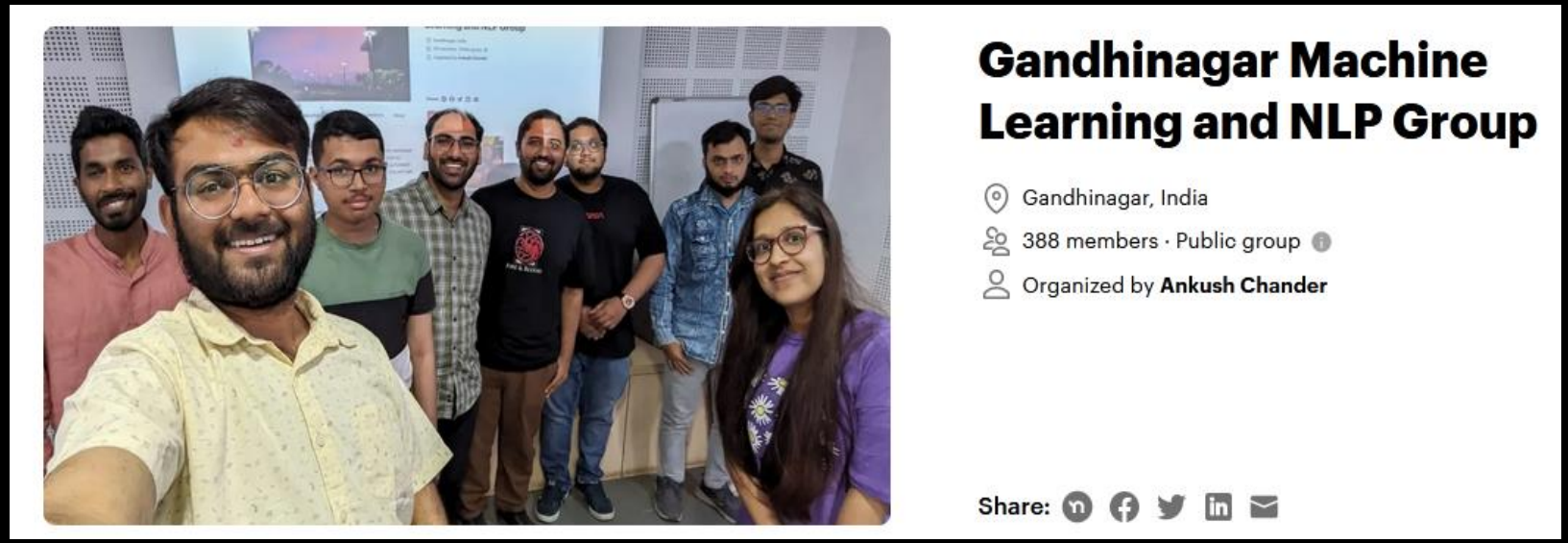


May 2024 Meetup!

Why Generate When You Can Discriminate? A Novel Technique for Text Classification using Language Models

Sachin Pawar, **Nitin Ramrakhiyani**, Anubhav Sinha, Manoj Apte, G.K. Palshikar

Published as Findings of the ACL (EACL 2024)

Introduction

- Text Classification: Task of assigning one or more applicable class labels to a given piece of text – sentence, paragraph, phrase, etc.

- Task: Use a causal / auto-regressive / decoder-only Language Model (LM) to perform text classification

- Some intuitive ideas that quickly come to our mind
  - Train / Fine-Tune a standard / neural classifier / transformer
  - Zero-shot: Construct a prompt and try on ChatGPT / GPT-3.5 / GPT-4
  - Few-shot: Include some examples in the prompt as part of In-Context Learning (ICL)

Motivation

- Train / Fine-Tune a standard / neural classifier / transformer
  - Requires training data, Specialized Hardware (in case of PLM FT)
  - Not robust to technical changes (number of classes, training data availability)
- Zero-shot: Construct a prompt and try on ChatGPT / GPT-3.5 / GPT-4
  - Privacy Concerns while using ChatGPT or large LMs
  - Local deployment preferred in commercial settings
  - Smaller models prone to hallucination when generating
- Few-shot: Include examples in the prompt as part of In-Context Learning (ICL)
  - Context window of the PLM restricts number of ICL examples.

Outline of the Proposed Approach

- **Step 1**: **Generating features using LM**
  - For any text X to be classified, elicit perplexity and log-likelihood features from LM using label-specific augmentations
  - `<X>. This text is about <key phrase>.`
  - E.g., `Indian stock markets gain on third consecutive day. This text is about economy.`
  - **Features** derived from - Conditional perplexity (and log-likelihood) of each key phrase given the text to be classified
    - Perplexity(`economy | Indian stock markets gain on third consecutive day. This text is about`)
- **Step 2**: **Learning a classifier**
  - Optionally, training a light-weight classifier like *Logistic Regression* or *Support Vector Machine* using training examples

Background
Revisiting some concepts

Key phrases

- We assume that each class can be represented using a small set of key phrases
  - E.g., **Business**: `business, stock market, banking, monetary investments, economy, income and expenditure, corporate profit and loss`
  - E.g., **MeanOfTransportation**: `a vehicle, a car, a train, an aeroplane, a ship or boat`
- Key phrases can be obtained from *domain experts* or from *documented domain knowledge* (e.g., audit checklist)
- Comparing with verbalizers used in prompt-tuning techniques like PET[1] or KPT[2]
  - These techniques are used with encoder-only models – whereas our technique is designed to work with decoder-only models
  - The verbalizers need to be single word only – whereas key phrases used by our technique can be multi-word

1. Schick and Schutze, Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference, EACL 2021
2. Hu et al., Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification, ACL 2022

Perplexity and Log-likelihood

- Perplexity is used as a metric to evaluate language models[3,4].
- We propose to use perplexity for a different purpose
  - judging plausibility of a text fragment using an autoregressive LM and
  - comparing multiple such text fragments to decide which one is the most plausible.

- By plausibility of a text, we mean that it is seemingly more reasonable or probable.
  - The second-in command of the Shiv Sena (UBT) is Sanjay Raut.

  is more plausible than
  - The second-in command of the Shiv Sena (UBT) is Nitin Ramrakhiyani. 

3. Jurafsky and Martin, Speech and Language Processing, 3rd Edition (Online)
4. Manning, Raghavan and Schutze, Introduction to Information Retrieval, Cambridge University Press, 2008

Perplexity and Log-likelihood

- Consider a text fragment $X = [w_1, w_2, \ldots, w_n]$. The perplexity of X as per a LM M is

$$PPL_M(X) = \prod_{i=1}^{n} \sqrt[n]{\frac{1}{P_M(w_i|w_{<i})}}$$

- The conditional perplexity of a text fragment X given another text $C = [c_1, c_2, \cdot\cdot, c_m]$ is

$$PPL_M(X|C) = \prod_{i=1}^{n} \sqrt[n]{\frac{1}{P_M(w_i|c_1, c_2, \cdots, c_m, w_{<i})}}$$

- Lower the perplexity, better is the plausibility of the text.

Perplexity and Log-likelihood

- Similarly, the log-likelihood and conditional log-likelihood for the fragment $X = [w_1, w_2, \ldots, w_n]$ (with $C = [c_1, c_2, \cdots, c_m]$) is

$$LL_M(X) = \sum_{i=1}^{n} log(P_M(w_i|w_{<i}))$$

$$LL_M(X|C) = \sum_{i=1}^{n} log(P_M(w_i|c_1, \cdots, c_m, w_{<i}))$$

- Higher the log-likelihood, better is the plausibility of the text.

- log-likelihood and perplexity

$$\begin{aligned} PPL_M &= \prod_{i=1}^{n} \sqrt[n]{\frac{1}{P_M(w_i|w<i)}} \\ &= \prod_{i=1}^{n} P_M(w_i|w<i)^{\frac{-1}{n}} \\ &= exp\{log \prod_{i=1}^{n} P_M(w_i|w<i)^{\frac{-1}{n}}\} \\ &= exp\{\frac{-1}{n} \sum_{i=1}^{n} log(P_M(w_i|w<i))\} \end{aligned}$$

Perplexity and Log-likelihood

- Similarly, the log-likelihood and conditional log-likelihood for the fragment $X = [w_1, w_2, \ldots, w_n]$ (with $C = [c_1, c_2, \cdots, c_m]$) is

$$LL_M(X) = \sum_{i=1}^{n} log(P_M(w_i|w_{<i}))$$

$$LL_M(X|C) = \sum_{i=1}^{n} log(P_M(w_i|c_1, \cdots, c_m, w_{<i}))$$

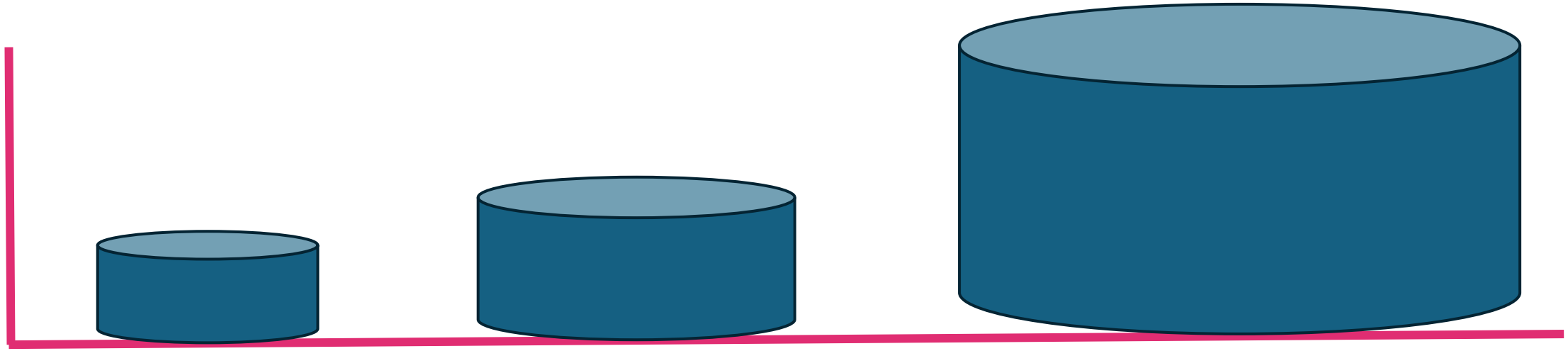
- Higher the log-likelihood, better is the plausibility of the text.

- log-likelihood and perplexity

$$\begin{aligned} PPL_M &= \prod_{i=1}^{n} \sqrt[n]{\frac{1}{P_M(w_i|w<i)}} \\ &= \prod_{i=1}^{n} P_M(w_i|w<i)^{\frac{-1}{n}} \\ &= exp\{log \prod_{i=1}^{n} P_M(w_i|w<i)^{\frac{-1}{n}}\} \\ &= exp\{\frac{-1}{n} \sum_{i=1}^{n} log(P_M(w_i|w<i))\} \end{aligned}$$

SLMs vs MedLMs vs LLMs

- Different sizing bounds being put in literature[5,6]



Small LMs: Less than 10B parameters

- Apple's OpenELM
- TinyLlama
- GPT-2, GPT2-XL, **GPTNeo-2.7B**
- GPT-J-6B,
- **Mistral-7B**, MPT-7B, Falcon-Instruct-7B, Falcon-40B

Medium LMs: 10B to 100B parameters

- GPT-NeoX-20B
- OPT-30B, Llama30B
- Falcon-40B

Large LMs: Greater than 100B parameters

- GPT-3.5, GPT-4
- Bloom
- PaLM

5. https://catherinebreslin.medium.com/small-language-models...
6. https://www.koyeb.com/blog/what-are-large-language-models

Approach Details

Formal Problem Definition

- **Input**:
  - $L = \{L_1, \cdots, L_C\}$ (set of C class labels)
  - $P_i = \{p_1^i, p_2^i, \cdots, p_{n_i}^i\}$ (set of n_i key phrases for each label $L_i \in L$)
  - X (text to be classified)
  - M (causal / decoder-only Language Model)
- **Output**:
  - One of more class labels ($\subset L$) which are assigned to X
- **Training regime**:
  - A small set of training examples where each instance is of the form $\langle X_t, L_t \rangle$
  - We consider at most 500 training instances

Step 1: Generating Features using LM (1/2)

- X : Text to be classified
- S : Connector sentence (e.g., `This text is about <key phrase>.`)
- For each class label L_i and for each of its key phrase p_j^i the following features are generated

$$f_{ij}^{PPL}(X) = \frac{PPL_M(p_j^i|X+S)}{PPL_M(p_j^i|S)}$$

$$\frac{PPL_M(economy|Indian\ stock\ markets\ gain\ on\ third\ consecutive\ day.This\ text\ is\ about)}{PPL_M(economy|This\ text\ is\ about)}$$

$$f_{ij}^{LL}(X) = LL_M\left(p_j^i|X+S\right) - LL_M\left(p_j^i|S\right)$$

Step 1: Generating Features using LM (1/2)

Text to be classified, $X =$ Expansion slows in Japan. Economic growth in Japan slows down as the country experiences a drop in domestic and corporate spending.	**Class labels with corresponding key phrases:** Sports: sports, a sporting event, a sportsperson, ... Business: business, economy, stock market, ... Science: science, space exploration, software, ...		
Label-specific augmentations of the above sentence		f_{ij}^{PPL}	f_{ij}^{LL}
A_{11}: Expansion slows in Japan. Economic growth in Japan slows down as the country experiences a drop in domestic and corporate spending. This news is about **sports**.		3.48	-2.50
A_{12}: Expansion slows in Japan. Economic growth in Japan slows down as the country experiences a drop in domestic and corporate spending. This news is about **a sporting event**.		1.42	-1.42
A_{21}: Expansion slows in Japan. Economic growth in Japan slows down as the country experiences a drop in domestic and corporate spending. This news is about **business**.		1.22	-0.40
A_{22}: Economic growth in Japan slows down as the country experiences a drop in domestic and corporate spending. This news is about **economy**.		0.62	0.95
A_{31}: Expansion slows in Japan. Economic growth in Japan slows down as the country experiences a drop in domestic and corporate spending. This news is about **science**.		7.12	-3.92
A_{32}: Expansion slows in Japan. Economic growth in Japan slows down as the country experiences a drop in domestic and corporate spending. This news is about **space exploration**.		1.52	-1.27

Step 1: Generating Features using LM (2/2)

- Class-level features are computed using key phrase level features

$$f_i^{PPL}(X) = \min_j f_{ij}^{PPL}(X)$$
$$f_i^{LL}(X) = \max_j f_{ij}^{LL}(X)$$

- **Zero-shot classification** is achieved by simply considering the ***minimum*** or ***maximum*** of these feature values

$$ZS\_PPL(X) = \operatorname{argmin}_i f_i^{PPL}(X)$$
$$ZS\_LL(X) = \operatorname{argmax}_i f_i^{LL}(X)$$

Step 2: Learning a Classifier

- This is an *optional* step and is necessary only in case of a *supervised setting*

- The perplexity and log-likelihood features are generated for all the training examples
- Horizontal scaling of these features to capture relative variations
- A light-weight classifier is trained using these features
  - Logistic Regression (LR)
  - Support Vector Machines (SVM)

- Naturally, there is *no limit on number of training examples* that can be used
  - We have used at most 500 examples in our experiments

Experimentation and Analysis

Datasets

Dataset	#instances		#labels	#key phrases
	train	test		
SST-2	500†	1821	2	20
TREC	500†	500	6	50
AGNews	500†	7600	4	37
DBPedia	500†	1000†	14	41
Ethos	200†	233†	8*	20

Baselines

- ZS-KP: Zero shot with description of key phrases of each class
- ZS-KP-CoT: Similar to ZS-KP with a Chain-of-Thought Prompt
- FS-ICL: In-context Learning using Few shots (k = 16)
- CHT: Classification Head Tuning (frozen LM)
- CHT-BERT: CHT over BERT CLS
- Min et al, Noisy Channel Language Model Prompting
  - “noisy channel” and "direct” methods to compute conditional probability of the input text given the label or vice versa (Few-shot ICL and prompt tuning)
- Estienne et al., Unsupervised Calibration through Prior Adaptation
  - Calibrate output probabilities of an LM through prior adaptation
  - Unsupervised (UCPA) where no labelled data is needed and
  - Semi-unsupervised (SUCPA) where some training examples (600) are used

Dataset	Label	Key phrases
SST-2	Positive	great, good, encouraging, brilliant, excellent, accurate, realistic, engaging, funny, exciting
	Negative	terrible, bad, unrealistic, frustrating, boring, forgettable, predictable, thoughtless, appalling, incomprehensible
AGNews	World	politics, terrorism, president of a country, a military related event, minister of a country, elections and government formation, a natural disaster, a war or an armed conflict, protests or demonstration, religious events
	Sports	sports, a sporting event, sporting awards, a sports champion, a sportsperson, wins or losses in sports, prize money
	Business	business, stock market, banking, monetary investments, economy, income and expenditure, corporate profit and loss, international trade, sale of goods and services, monetary policies
	Science	science, technology and engineering, research and development, internet and web, space exploration, cyber security, software, weather and climate, healthcare and pharma, flora and fauna
TREC	ABBR	an abbreviation, an expression which is abbreviated
	ENTY	an entity, an animal, an organ of body, a color, an invention, book and other creative piece, a currency name, a disease or a medicine, an event, food, a musical instrument, a language, a letter or a character, a plant, a product, a religion, a sport , a chemical element or a substance , a symbol or a sign, a technique or a method, an equivalent term, a vehicle, a word with a special property
	DESC	description of something, a definition of something, a manner of an action, a reason
	HUM	an individual, a group or organization of persons, a title of a person, description of a person

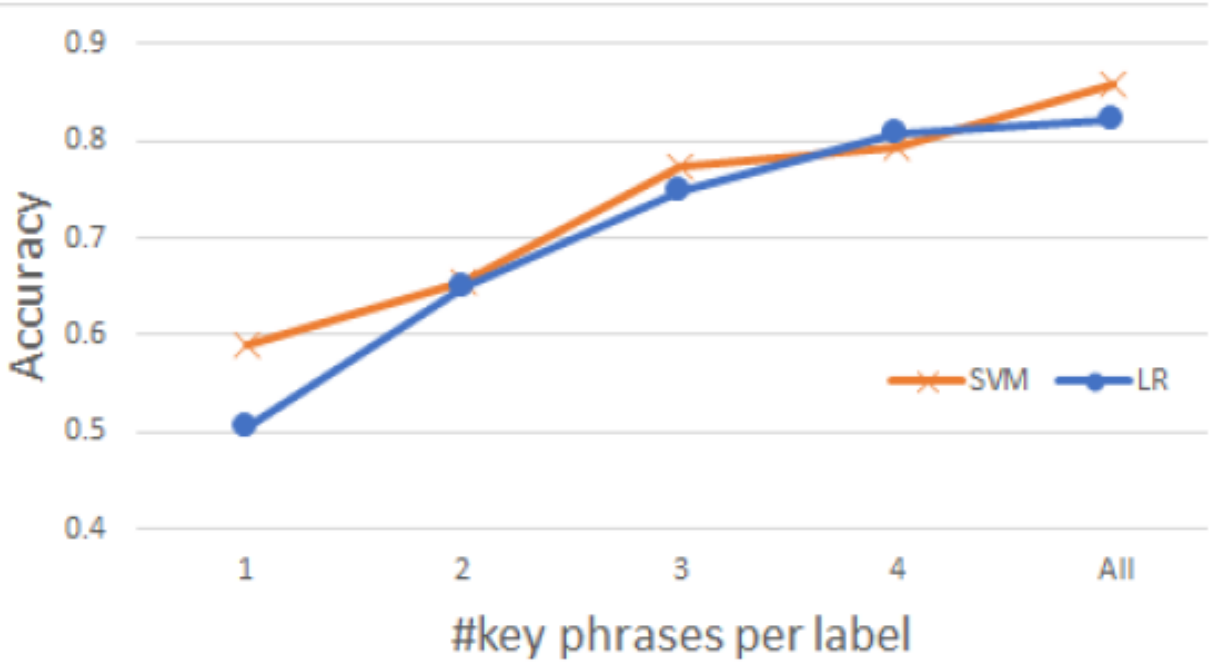
Comparison with Baselines using GPT2-XL Model

	SST-2	TREC	AGNews	DBPedia	Ethos
Unsupervised Calibration through Prior Adaptation (Estienne, 2023)					
SUCPA (zero-shot)	0.850	0.460	0.700	0.660	NA
SUCPA (few-shot)	0.890	0.550	0.780	0.880	NA
Noisy Channel Language Model Prompting† (Min et al., 2022)					
Channel (zero-shot)	0.771	0.305	0.618	0.514	NA
Channel (concat-based)	0.850	0.420	0.685	0.585	NA
Channel (ensemble-based)	0.775	0.315	0.743	0.648	NA
Other baselines:					
ZS-KP (zero-shot with keyphrases)	0.183	0.10	0.088	0.157	0.137
ZS-KP-CoT	0.160	0.01	0.029	0.089	0.032
FS-ICL	0.874	0.476	0.330	0.085	0.182
CHT	0.567	0.476	0.592	0.488	0.029
CHT-BERT*	0.890	0.698	0.801	0.834	0.219
Our proposed techniques:					
ZS-PPL (zero-shot with only PPL features)	0.871	0.478	0.776	0.762	0.479
ZS-LL (zero-shot with only LL features)	0.875	0.462	0.764	0.716	0.421
SVM with all features and horizontal scaling	0.919	**0.860**	0.851	0.912	0.707
LR with all features and horizontal scaling	**0.920**	0.824	**0.853**	**0.924**	**0.715**

Comparison using the GPT-Neo-2.7B model

	SST-2	TREC	AGNews	DBPedia	Ethos
Baselines:					
ZS-KP (zero-shot with keyphrases)	0.248	0.020	0.039	0.182	0.035
ZS-KP-CoT (ZS-KP with Chain-of-Thought)	0.061	0.046	0.024	0.239	0.019
FS-ICL	0.814	0.308	0.672	0.689	0.438
CHT	0.620	0.734	0.691	0.558	0.164
Our proposed techniques:					
ZS-PPL (zero-shot with only PPL features)	0.752	0.384	0.787	0.735	0.527
ZS-LL (zero-shot with only LL features)	0.766	0.418	0.774	0.67	0.438
SVM with all features and horizontal scaling	**0.893**	**0.804**	**0.860**	0.912	0.671
LR with all features and horizontal scaling	**0.893**	0.798	0.858	**0.926**	**0.673**

Effect of varying number of key phrases and training examples (TREC)



Ablation and Effect of Connector Sentences

	SST-2	TREC	AGNews	DBPedia	Ethos
LR default setting: With all features and horizontal scaling	**0.920**	**0.824**	0.853	**0.924**	**0.715**
LR default setting without Horizontal scaling	0.908	0.820	0.792	0.911	0.686
LR default setting without LL features	0.914	0.684	0.828	0.884	0.633
LR default setting without PPL features	0.919	**0.824**	**0.856**	0.916	0.712
LR default setting without class-level features	0.918	0.822	0.850	0.917	0.703
LR default setting without keyphrase-level features	0.880	0.486	0.784	0.886	0.672
LR default setting with only one keyphrase per class	0.832	0.504	0.688	0.855	0.647

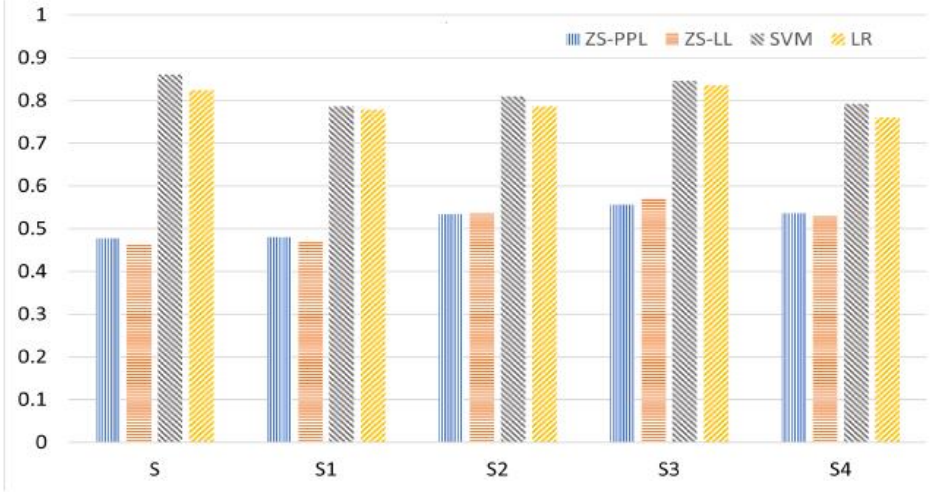


Figure 3: Accuracy for TREC with various connector sentences using GPT2-XL (S:`The answer will be`, S1: `The answer will be about`, S2: `The answer is`, S3:`The answer must be`, S4: `The answer is about`)

Explainability through Key Phrases (Example from AGNews)

Text	Label	Key phrase	f_{ij}^{PPL}
Afghan Army Dispatched to Calm Violence. KABUL, Afghanistan - Government troops intervened in Afghanistan's latest outbreak of deadly fighting between warlords, flying from the capital to the far west on U.S. and NATO airplanes to retake an air base contested in the violence, officials said Sunday...	World	terrorism	0.259
Late rally sees Wall Street end week on a positive note. US BLUE-chips recovered from an early fall to end higher as a drop in oil prices offset a profit warning from aluminium maker Alcoa, while a rise in Oracle fuelled a rally in technology stocks after a judge rejected a government attempt to block a...	Business	stock market	0.087
Bekele, Isinbayeva top track athletes. Names Ethiopian distance runner Kenenisa Bekele and Russian pole vaulter Yelena Isinbayeva were named male and female athletes of the year by the world track and field federation. Isinbayeva set eight world records in 2004, including one while winning the gold medal at the Olympics. Bekele won the 10,000 meters in Athens and finished second to Hicham El Guerrouj in ...	Sports	sporting awards	0.072
Plans for new Beagle trip to Mars. The team behind Beagle 2, the failed mission to land on Mars and search for life, have unveiled plans for a successor. Professor Colin Pillinger, lead...	Science	space exploration	0.183

Real-life Application: Analysis of Financial Audit Reports

- **Use-case**: To automatically check whether all important Audit Aspects are being covered in a financial audit report
- **Text Classification Task**: To classify each sentence in Audit Report in *one or more of 15 Audit Aspects*
  - Payables, Receivables, Inventory, Fixed Assets, etc.
- **Training data**: Silver standard annotations using a set of regular expressions
- **Test data**: 10 Audit reports manually annotated (*1668 sentences*)

Table: micro-averaged F1-scores on the test dataset, using GPT2-XL

#training instances	SVM	LR	ZS-PPL	ZS-LL	ChatGPT
1097	0.542	0.536	0.380	0.410	0.520
500	0.503	0.498			

Conclusions and Future Work

- Proposed a two-step technique for text classification using moderate-sized causal Language Models
- A new way of exploiting available training examples in addition to in-context learning and fine-tuning
  - No limit on number of training examples
  - No need for parameter updates
  - Explainability through most suitable key phrases
  - Applicability in resource poor environments

- In future, we plan to extend to this work by:
  - Automatically discovering optimal key phrases and connector sentences
  - Exploring ensemble of multiple small LMs

Thank you for listening!

Questions? | Thoughts | Discussion

Have more questions? Send them to
nitin.ramrakhiyani@gmail.com