Multilingual GLM

孙梦阳，李天健，杜政晓

Outline

- Motivation
- MultiGLM
- Prompt Based Fine Tuning
- Experiments
- Limitations & Future Directions

Motivation

- Impractical to train one model for each language. Want to perform cross-lingual tasks (Machine Translation, Summarization)
- Impractical to train one model for each task. Want to perform tasks on low resource languages by transfer learning.
- In depth linguistics studies concentrate on English(semantic, syntactic). Want to facilitate multilingual research and preserve endangered languages.

Motivation - Related Studies

- mBERT (Devlin et al., NAACL 2019)
  - Multilingual variant of BERT
  - 180M parameters, 104 languages
- XLM-R (Conneau et al., NeurIPS 2019)
  - Multilingual variant of RoBERTa(Liu et al., 2019)
  - 550M parameters, 100 languages
- mT5 (Xue et al., NAACL 2021)
  - Multilingual Variant of T5(Raffel et al, JMLR 2020)
  - Up to 13B parameters, 104 languages
  - Achieved SOTA on sentence pair, NER and QA benchmarks

- MultiGLM
  - Multilingual variant of GLM(Du et al., ACL 2022)
  - 1B parameters, 104 languages
  - In the case where the parameters (1B V.S. 1.2B) and the number of training tokens (643B V.S. 1T) are both smaller than mT5, outperform the latter on question answering tasks

Xue et al., (2021) mT5: A massively multilingual pre-trained text-to-text transformer

Outline

- Motivation
- MultiGLM
- Prompt Based Fine Tuning
- Experiments
- Takeaways

GLM - Autoregressive Pretraining

The quick brown fox jumps over the lazy dog

Autoregressive: GPT-2 (Radford et al., 2019)

The quick brown fox jumps [gMASK]

Autoencoding: BERT (Devlin et al., 2018)

[CLS] The quick [MASK] fox jumps over the lazy dog

GLM - Autoregressive Pretraining

Type	Autoregressive	Autoencoding
Example	GPT-2(Radford et al., 2019)	BERT(Devlin et al., 2018)
Independence Assumption	No assumption	Masked tokens are independent
Input Noise	No input noise	Task discrepancy between pretrain and finetune
Context Dependency	Unidirectional	Bidirectional

Table 1: Comparison between autoregressive and autoencoding language models

Du et al., (2022) GLM: General Language Model Pretraining with Autoregressive Blank Infilling

GLM - Autoregressive Pretraining

The quick brown fox jumps over the lazy dog

Autoregressive Blank Infilling: GLM

The [sMASK$_1$] fox jumps [sMASK$_2$] the [sMASK$_3$]

The [sMASK$_2$] fox jumps [sMASK$_3$] the [sMASK$_1$]

The [sMASK$_3$] fox jumps [sMASK$_1$] the [sMASK$_2$]

(Train over all 6 permutations)

Du et al., (2022) GLM: General Language Model Pretraining with Autoregressive Blank Infilling

GLM - Autoregressive Pretraining

The quick brown fox jumps over the lazy dog

Autoregressive Blank Infilling: GLM

The [sMASK$_1$] fox jumps [sMASK$_2$] the [sMASK$_3$]

Assume Model predicts [sMASK$_1$] = witty brown

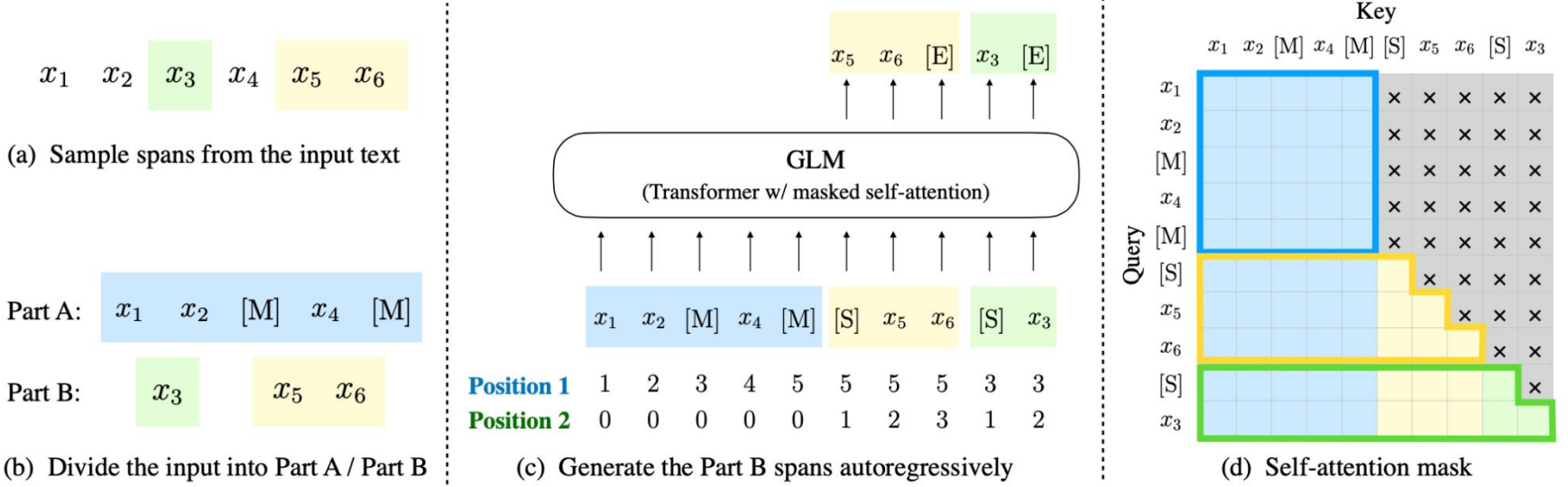
- [sMASK$_2$] = $\max_Y P(Y|\text{The witty brown fox jumps, the})$

Assume Model predicts [sMASK$_2$] = over

- [sMASK$_3$] = $\max_Y P(Y|\text{The witty brown fox jumps over the})$

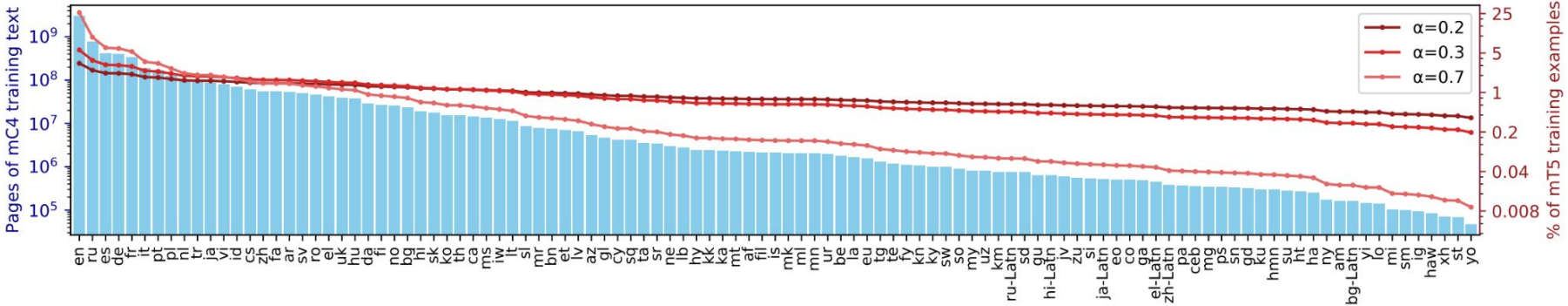
Du et al., (2022) GLM: General Language Model Pretraining with Autoregressive Blank Infilling

GLM - Autoregressive Pretraining



Du et al., (2022) GLM: General Language Model Pretraining with Autoregressive Blank Infilling

MultiGLM - “Curse of Multilinguality”



High Resource:
EN: 1.5T
FR: 200G
DE: 400G
ES: 300G

Low Resource
SI: 1G
HA: 80M
SU: 60M

α = 1 means no sampling

High α(0.7) - Better Performance on High Resource languages

Low α(0.2) - Better Performance on Low Resource languages

Xue et al., (2021) mT5: A massively multilingual pre-trained text-to-text transformer

Outline

- Motivation
- MultiGLM
- Prompt Based Fine Tuning
- Experiments
- Limitations & Future Directions

Prompt Based Fine Tuning - Pattern & Verbalizer

- Pattern

(QA) **`Question`**: Your Question Here**`, Answer`**: [MASK] **`.`**
(Sentiment) Your sentence Here **`, It was`** [MASK] **`.`**
(NLU) Sentence 2 **`, True or False?`** [MASK] **`.`** Sentence 1 **`.`**

- Verbalizer

(QA) No Verbalizer
(Sentiment) Awful - 0, Average - 1, Great - 2, Wonderful - 3
(NLU) True - 1, False - 0

Schick and Schütze(2020) Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference

Prompt Based Fine Tuning - Multilingual Prompting

- Directly Translating the Template For the Task Language

  (Zhao and Schütze EMNLP 2021)
- One training example, multilingual templates (Qi et al., ACL 2022)
- Universal template + multilingual verbalizers (Zhou et al., ACL 2022)

Zhou et al.,(2022) Enhancing Cross-lingual Prompting with Mask Token Augmentation

Direct Translation during Testing

- Directly translating the template for the task language (Zhao and Schütze EMNLP 2021)
  - Comparable but still not better than English only templates.

English(Training):

Today is a sunny day. `Question`: Is it a good day to go out ? `Answer`: [MASK] .

中文(Testing):

今天是个晴天。问题:今天适合出去玩吗？答案:[MASK]。

Français(Testing):

Aujourd'hui est une journée ensoleillée. `Question`： Est-ce qu'aujourd'hui est un bon jour pour sortir ? `Réponse`: [MASK].

Zhao and Schtüze(2021) Discrete and Soft Prompting for Multilingual Models

Cross-lingual Templates

- One training example, multilingual templates (Qi et al., ACL 2022)
  - Only Experimented on XNLI Sentence Pair Task

English(Training):

Today is a sunny day. `Question`: Is it a good day to go out ? `Answer`: [MASK] .

中文(Training):

Today is a sunny day。问题:Is it a good day to go out？答案:[MASK]。

Français(Training):

Today is a sunny day . `Question`： Is it a good day to go out? `Réponse`: [MASK].

Qi et al.,(2022) Enhancing Cross-lingual Natural Language Inference by Prompt Learning from Cross-lingual Templates

Universal Template + Multilingual Verbalizers

Universal template + Multilingual Verbalizers (Zhou et al., ACL2022)

Language	Verbalizer
EN	Paraphrase → yes Non-paraphrase → no
DE	Paraphrase → Ja Non-paraphrase → Nein
ES	Paraphrase → sí Non-paraphrase → no
FR	Paraphrase → Oui Non-paraphrase → non
JA	Paraphrase → はい Non-paraphrase → ない
ZH	Paraphrase → 是 Non-paraphrase → 否
KO	Paraphrase → 예 Non-paraphrase → 아니

Table 5: The multilingual verbalizer for PAWS-X.

English:

Today is a sunny day. `Question`: Is it a good day to go out ?

`Answer`: [MASK] .

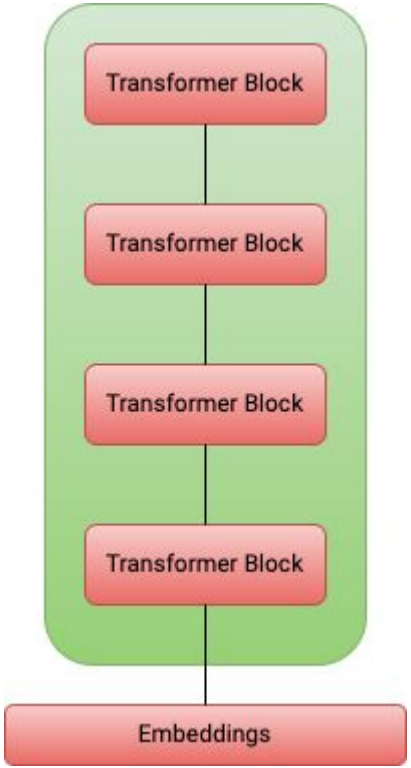
Universal :

Today is a sunny day. Is it a good day to go out? [MASK]

A. B? [MASK]

Zhou et al.,(2022) Enhancing Cross-lingual Prompting with Mask Token Augmentation

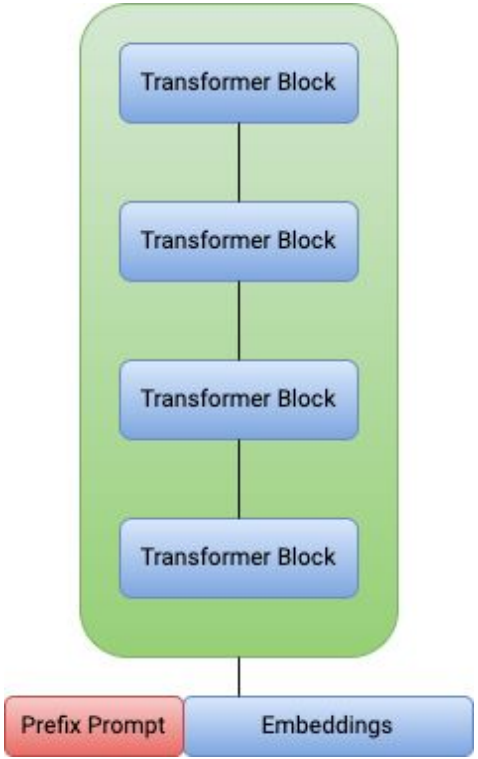
Prompt Based Fine Tuning V.S. Prompt Tuning



- Standard & Prompt Based Fine Tuning
  - Updates **ALL** the parameters of the model
  - Faster Convergence
  - Space Consuming

    (For 1B model, checkpoint approx. 12-14GB)
  - Best performance, Especially for smaller models(under 10B)

Prompt Based Fine Tuning V.S. Prompt Tuning

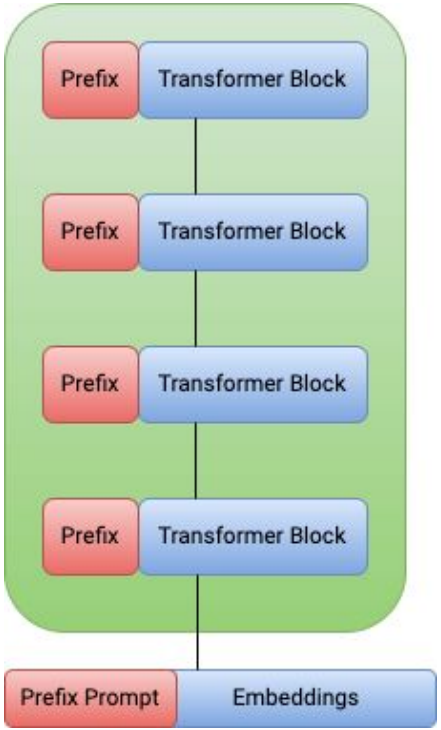


- Prompt Tuning
  - Freezes **ALL** the parameters of the model
  - Only Finetune Prefix Embeddings
  - Slower Convergence
  - Significant less space(approx 0.1% of parameters)
  - Comparable Performance for Large Models

Lester et al.,(2021) The Power of Scale for Parameter-Efficient Prompt Tuning

Liu et al.,(2021) GPT Understands, too

Prompt Based Fine Tuning V.S. Prompt Tuning



- **Deep** Prompt Tuning
  - Freezes **MOST** parameters of the model
  - Finetune Prefix Embeddings and Prefix in Model
  - Better Performance than “shallow” prompt tuning.

Li and Liang(2021) Prefix-Tuning: Optimizing Continuous Prompts for Generation

Qin and Eisner(2021) Learning How to Ask: Querying LMs with Mixture of Soft Prompts

Liu et al.,(2021) P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks

Outline

- Motivation
- MultiGLM
- Prompt Based Fine Tuning
- Experiments
- Limitations & Future Directions

Experiments - Zero Shot Transfer

Model	Params	XNLI	PAWS-X	XQuAD	MLQA	TyDiQA
		Acc.	Acc.	F1/EM	F1/EM	F1/EM
mBERT	180M	65.4	81.9	64.5/49.4	61.4/44.2	59.7/43.9
XLM-R	550M	79.2	86.4	76.6/60.8	71.6/53.2	65.1/45.0
MMTE	565M	67.4	81.3	64.4/46.2	60.3/41.4	43.6/29.1
mT5-Small	300M	67.5	82.4	58.1/42.5	54.6/37.1	35.2/23.2
mT5-Base	580M	75.4	86.4	67.0/49.0	64.6/45.0	57.2/41.2
mT5-Large	1.2B	81.1	88.9	77.8/61.5	71.2/51.7	69.9/52.2
mT5-XL	3.7B	82.9	89.6	79.5/63.6	73.5/54.5	75.9/59.4
mT5-XXL	13B	85.0	90.0	82.5/66.8	76.0/57.4	80.8/65.9
mGLM-Large	1B	74.7*	85.8*	**83.6/71.9**	67.4/**54.3**	67.1/**53.6**

Table 2: experiment resutls, * indicates that the result is evaluated on validation set instead of test set.

Experiments - Zero Shot Transfer

Prompt	F1/EM
English	**67.52/54.34**
Multilingual	67.39/54.2
Universal	59.4/45.2

Results with different prompting methods on MLQA dataset

Setup	F1/EM
Zero-shot/Full	**68.0/53.87**
Zero-shot/P-Tuning	51.6/40.2
Gold data/Full	81.8/71.6
Gold data/P-Tuning	79.5/69.7

Results with different fine tuning methods on TyDiQA dataset

Lester et al.,(2021) The Power of Scale for Parameter-Efficient Prompt Tuning

Experiments - Cross-lingual Text Summarization

Authorities in Shanghai have announced that some Covid-19 lockdown measures imposed on businesses will be lifted from Wednesday. Plans have also been introduced to support the city's economy, which has been hit hard by the restrictions. The commercial centre has been under a strict lockdown for almost two months. Meanwhile, China's capital Beijing has reopened parts of its public transport system as well as some shopping malls and other venues as infections ease. The announcement in Shanghai came as official figures showed on Sunday that new daily coronavirus cases fell to 122 from 170 over the previous 24 hours. Officials said guidelines to curb the spread of Covid-19 and control the number of people returning to work will be revised. The move will see "unreasonable restrictions" being lifted on restarting work and production at companies, vice mayor Wu Qing told a news briefing.

中文摘要:上海将于周三解除对企业实施的封锁措施。北京重新开放部分交通系统。

本田圭佑(ほんだ けいすけ、1986年6月13日 -)は、日本のサッカー選手、サッカー指導者、実業家。ポジションはミッドフィールダー。元日本代表。カンボジア代表のGM・監督を務める。アジア(日本)、ヨーロッパ(オランダ・ロシア・イタリア・アゼルバイジャン・リトアニア)、北中米(メキシコ)、オセアニア(オーストラリア)、南米(ブラジル)の5地域のプロリーグでプレーして得点を決めた。日本代表時代には日本人初となるW杯3大会連続ゴールを決め(日本人のW杯最年長得点者)、アジア人のW杯最多得点者となり、アジア人初となるW杯3大会連続アシストも達成。ルディ・フェラー、デイビッド・ベッカムらに続くW杯3大会連続となる得点とアシストの両方を記録した史上6人目の選手となった。FIFA選出のW杯の「マン・オブ・ザ・マッチ」数において日本人最多の4回」受賞し、ミロスラフ・クローゼ、エデン・アザール、ハメス・ロドリゲスらと並び受賞総数5位である。2020年にイギリスメディア『90min』選定の「21世紀の日本代表ベストイレブン」に選出、ドイツメディア『SPOX』選定の「アジア歴代ベストイレブン」に選出、『アジアサッカー連盟(AFC)』『Opta』選定の「W杯アジアベストイレブン」に選出された。現役選手ながら選手以外の活動も精力的に行い、2012年から自身がプロデュースするサッカースクール「SOLTILO FAMILIA」を日本全国で開校。2015年からは複数のプロサッカークラブの実質的なオーナーを務め、2018年にはカンボジア代表GM・監督に就任した。

中文摘要:本田圭佑是日本足球运动员、教练和实业家。1986年6月13日出生于日本。

Experiments - Multilingual Factual Probing

Language Models as Knowledge Bases(w/o finetune)

问题:中国的首都是哪里?
答案:[中国首都北京。]

Question: Where is the city of Ann Arbor located?
Answer: [Michigan.]

質問:日本の国花は何ですか?　　(日本的国花是什么)
答え:[日本の国花は、桜です。]　　(日本的国花是樱花)

Petroni et al.,(2019) Language Models as Knowledge Bases?

Outline

- Motivation
- MultiGLM
- Prompt Based Fine Tuning
- Transfer Learning Experiments & Demos
- Limitations & Future Directions

Limitations - Academic Article Summarization

- Failing to recognize terminologies (Abstract of Vaswani et al., 2017)

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 Englishto-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

中文摘要:我们提出了一种新的简单网络架构,即变压器。模型在两项机器翻译任务中都取得了优异的成绩。

Limitations - Academic Article Summarization

- Failing to recognize terminologies (Abstract of Zhou et al., 2022)

  Prompting shows promising results in few- shot scenarios. However, its strength for multilingual/cross-lingual problems has not been fully exploited. Zhao and Schütze (2021) made initial explorations in this direction by presenting that cross-lingual prompting outperforms cross-lingual fine-tuning. In this paper, we conduct empirical analysis on the effect of each component in cross-lingual prompting and derive Universal Prompting across languages, which helps alleviate the discrepancies between source-language training and target-language inference. Based on this, we propose a mask token augmentation framework to further improve the performance of prompt-based cross-lingual transfer. Notably, for XNLI, our method achieves 46.54% with only 16 English training examples per class, significantly better than 34.99% of fine-tuning.

  中文摘要: Zhao和Schütze提出了跨语言发短信的影响。他们提出了一种增强面具身份的框架。

Future Directions - Retrieval Augmented LMs

	Retriever	Decoder	Training Objective	Remarks
GraphQA	Graph Based	MLP	Marginal Likelihood	Vertices = Passages Edges = BM25 Similarity
ORQA(ACL '19)	BERT[CLS]	BERT Concat	Inverse Cloze Task(ICT)	Softmax operation computationally expensive
REALM(ICML '20)	BERT[CLS]	BERT Concat	Masked Language Model	Document encoder updated asynchronously
DPR(EMNLP '20)	BERT[CLS]	BERT Multiply	Neg. log prob. of positive documents	Only the retriever is trained Question = Gold Data
RAG(NIPS '20)	BERT[CLS]	BART AR*	Marginal Likelihood	Document encoder not updated Seq2Seq autoregressive decoder
ProQA(EACL '21)	BERT[CLS]	BERT	Neg. log prob. of positive documents + answer scores	Questions = BART generations Trained end-to-end. Treats all retrieved docs as a huge doc.

Table 1: Comparison between different retrieval augmented LMs, *Autoregressive

Thank you !