BLADE: Enhancing Black-box Large Language Models with Small Domain-Specific Models

разбор статьи
https://arxiv.org/abs/2403.18365

Дмитрий Колодезев @promsoft kolodezev.ru
2024.04.09 @ DS TALKS

LLM are capable, however...

BLADE: Enhancing Black-box Large Language Models with Small Domain-Specific Models

Haitao Li
DCST, Tsinghua University
Quan Cheng Laboratory
liht22@mails.tsinghua.edu.cn

Qingyao Ai*
DCST, Tsinghua University
Quan Cheng Laboratory
aiqy@tsinghua.edu.cn

Jia Chen
DCST, Tsinghua University
Quan Cheng Laboratory
chenjia0831@gmail.com

Qian Dong
DCST, Tsinghua University
Quan Cheng Laboratory
dq22@mails.tsinghua.edu.cn

Zhijing Wu
Beijing Institute of Technology
zhijingwu.bit.edu.cn

Yiqun Liu
DCST, Tsinghua University
Zhongguancun Laboratory
yiqunliu@tsinghua.edu.cn

Chong Chen
Huawei Cloud BU
chenchong55@huawei.com

Qi Tian
Huawei Cloud BU
tian.qi1@huawei.com

ABSTRACT

Large Language Models (LLMs) like ChatGPT and GPT-4 are versatile and capable of addressing a diverse range of tasks. However,

KEYWORDS

Large Language Models, Domain Adaptation, Bayesian Optimization

中国人发表了许多科学文章

- Я еще не весь китайский выучил,
  но там написано, что китайцы много публикуются

清華大學 计算机科学与技术系
Department of Computer Science and Technology, Tsinghua University

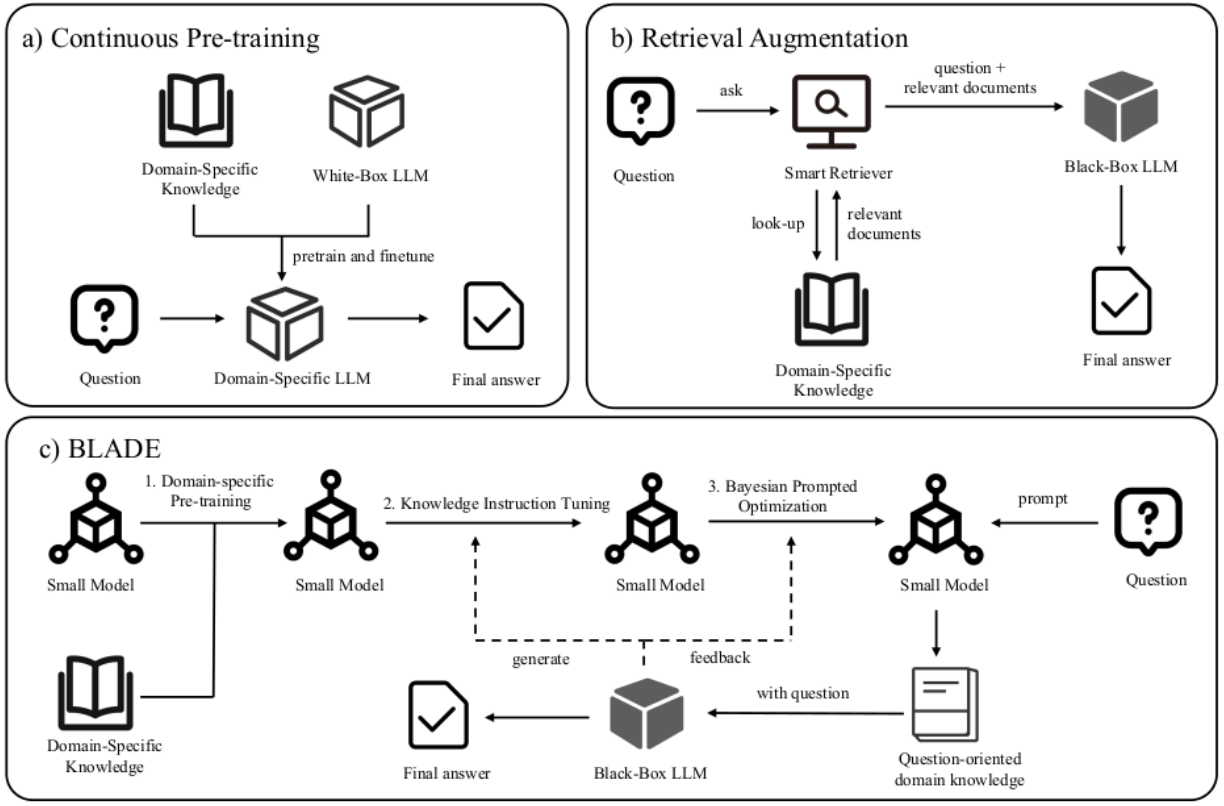
泉城实验室
Quan Cheng Laboratory

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

HUAWEI

Кибербезопасность?
DELTA Legal Case Retrieval
Те же авторы и данные

Третий путь


a) Continuous Pre-training
Domain-Specific Knowledge
White-Box LLM
pretrain and finetune
Question
Domain-Specific LLM
Final answer
b) Retrieval Augmentation
Question
ask
Smart Retriever
question + relevant documents
Black-Box LLM
look-up
relevant documents
Domain-Specific Knowledge
Final answer
c) BLADE
Small Model
1. Domain-specific Pre-training
Small Model
2. Knowledge Instruction Tuning
Small Model
3. Bayesian Prompted Optimization
Small Model
prompt
Question
generate
feedback
Domain-Specific Knowledge
Final answer
Black-Box LLM
with question
Question-oriented domain knowledge

Мы команда

- Большая LLM «общего назначения»,
  - Черный ящик
  - Не учим
  - Отвечает за «мышление» в целом
- «Маленькая» специализированная LLM
  - Менее 3b параметров
  - Учим по специальному протоколу
  - Отвечает за знания доменной области

В чем сила

- Маленькая LLM — маленькая, удобно учить
- Отделяем запоминание фактов от генерации
- В отличие от тупой близости по эмбеддингам в RAG у нас острый token-level cross attention
- Быстро
- Дешево

По шагам

- Domain-specific Pre-training
  - Вносим доменные знания в маленькую LLM
- Knowledge Instruction Tuning
  - Учит маленькую LLM следовать инструкциям
- Bayesian Prompted Optimization
  - Учит маленькую LLM говорить, чтобы большая понимала

Domain-specific Pre-training

- Просто учим маленькую модель генерировать тексты предметной области, которые ей показали

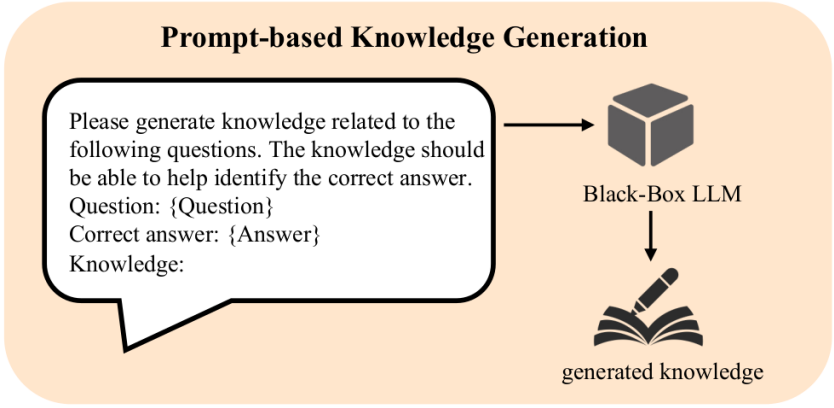
cally, given domain-specific unsupervised text $T = \{t_1, t_2, ..., t_n\}$, we optimize the model by maximizing the following training objective:

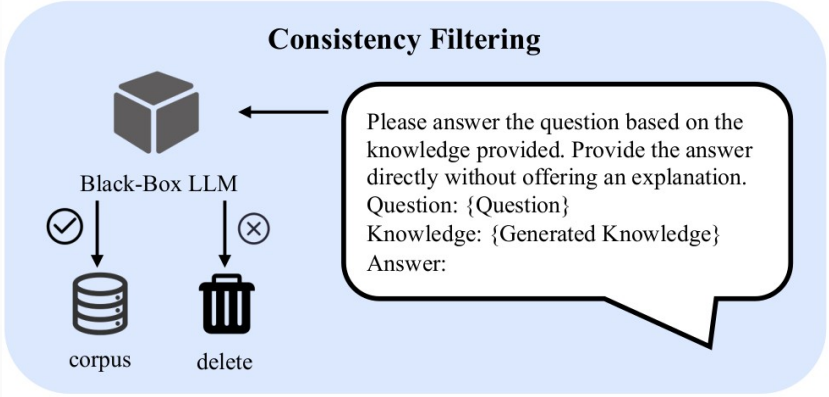
$$G(T) = \sum_i \log P\left(t_i \mid t_{i-k}, \ldots, t_{i-1}; \Theta\right),$$

where Θ is the parameter of the model. P is the conditional probability of generating the current token based on the previous tokens.

Knowledge Instruction Tuning

- Prompt-based Knowledge Generation
- Consistency Filtering
- Instruction Tuning





Bayesian Prompted Optimization

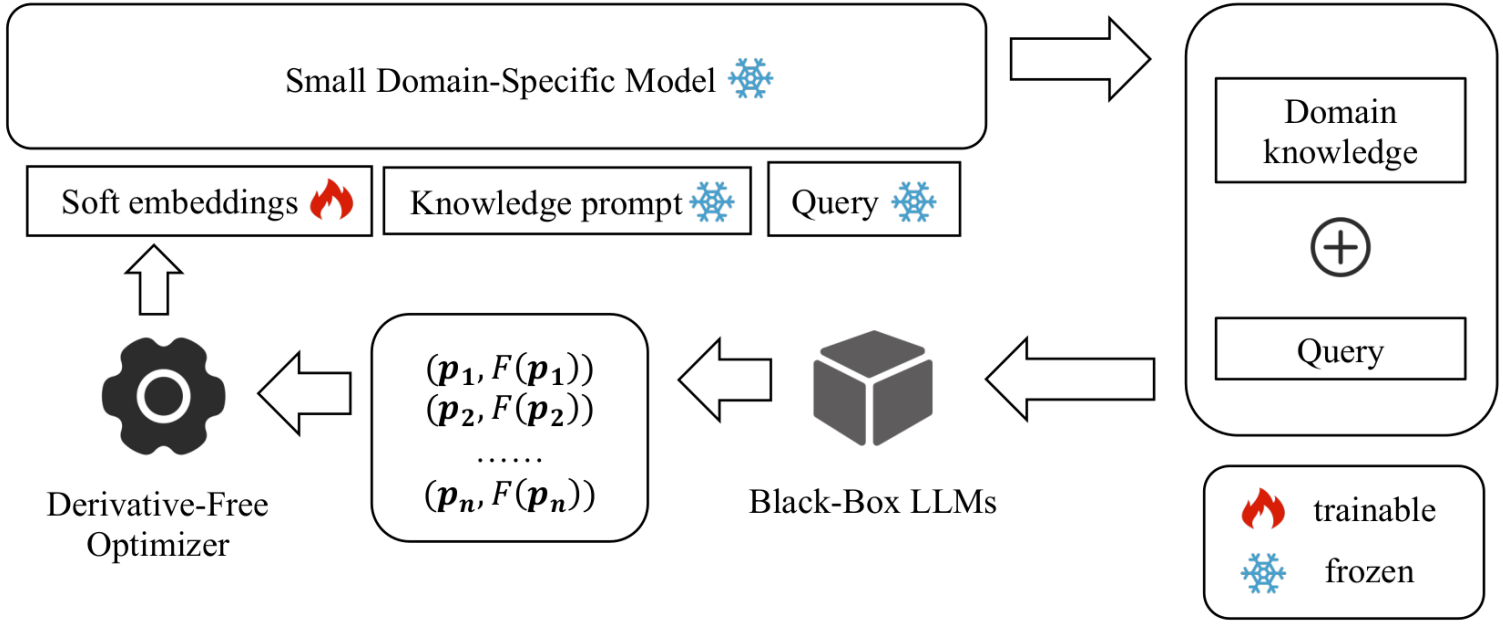


Figure 3: Illustration of the Bayesian Prompted Optimization where only soft embeddings are trainable. $F(p)$ is the objective score corresponding to soft embedding p. In each iteration, the derivative-free optimizer explores new soft embedding based on previous evaluation scores. The knowledge prompt is consistent with the instruction used in the Prompt-based Knowledge Generation stage.

BPO поподробнее

- Оптимизируем из конца в конец обе модели
- Большая — черный ящик, градиенты не ходят
- P-tuning с помощью derivative-free optimization
- Трудно! Но внутренняя размерность LLM мала
- Построим случайную проекцию d « D
- Сохраняет дистанцию (~) так что «близость» консистентна между пространствами

BPO поподробнее

- Суррогатная модель $F(\boldsymbol{p}) \sim GP(\mu, \sigma^2)$
- Функция выборки - Expected Improvement (EI)
- $$\boldsymbol{p_{n+1}} \in \arg\max_{\boldsymbol{p}\in\mathbb{R}^d} \mathbb{E}_{F(\boldsymbol{p})\sim GP(\mu,\sigma^2)} \left[\max\left\{0, F(\boldsymbol{p}) - \max_{i\in[n]} F(\boldsymbol{p}_i)\right\}\right],$$
- Картинки из другой статьи!

input : input domain $\mathcal{X}$, dataset $\mathcal{D}$, GP prior $\mathcal{M}_0 = \mathcal{GP}(\mu_0, k_0)$,
acquisition function α, noise ε

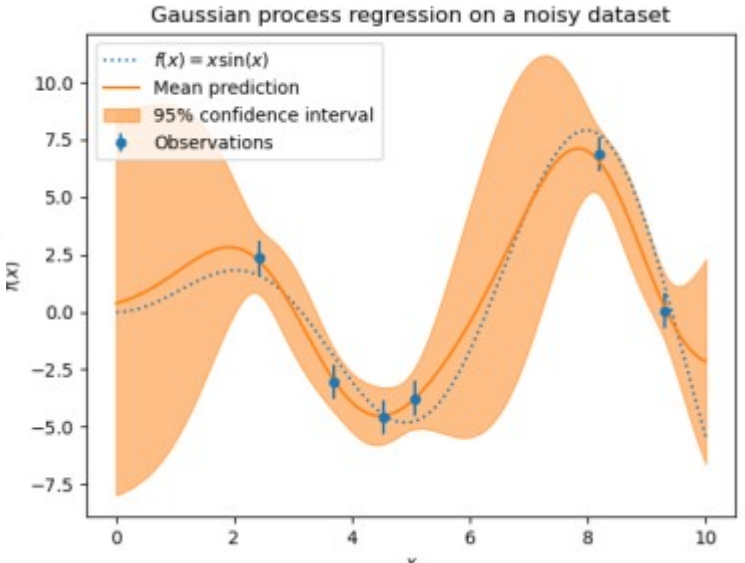
for $i = 1, 2, 3, \ldots$ **do**

$x_i \leftarrow \text{argmin}_{x\in\mathcal{X}} \alpha(x|\mathcal{M}_{i-1})$; // optimize α

$y_i \leftarrow f(x_i) + \varepsilon$; // do observation

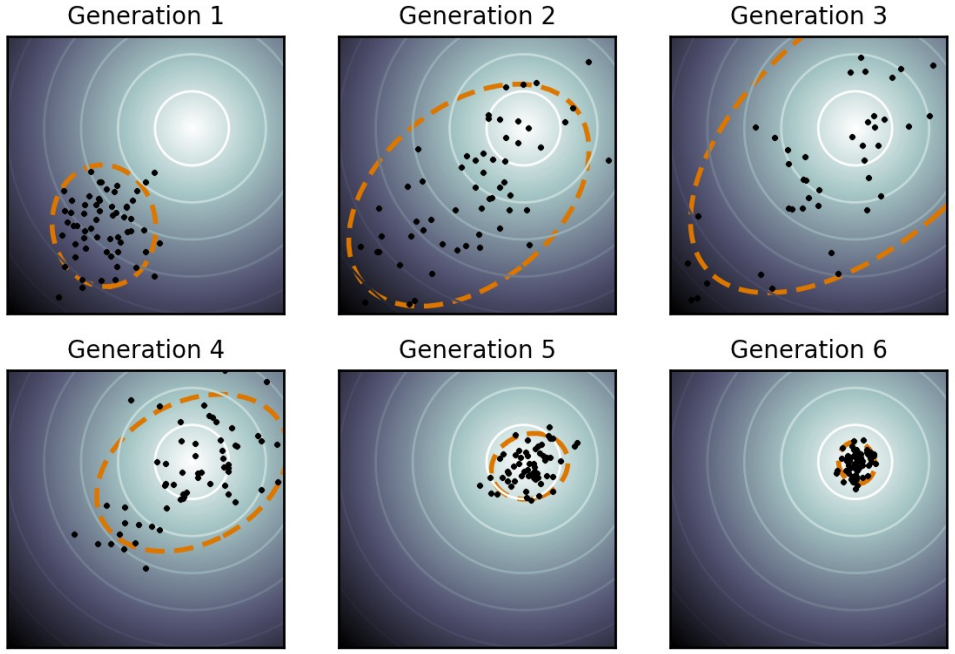
$\mathcal{M}_i \leftarrow \mathcal{M}_{i-1}|(x_i, y_i)$; // update model

end



BPO: CMA-ES

- Max-likelihood
- «Инерция»
- Похоже на Adam
- Но без градиента



https://en.wikipedia.org/wiki/CMA-ES

CMA-ES

set λ // number of samples per iteration, at least two, generally > 4
initialize m, σ, $C = I$, $p_\sigma = 0$, $p_c = 0$ // initialize state variables
while *not terminate* **do** // iterate
    for i **in** $\{1 \dots \lambda\}$ **do** // sample λ new solutions and evaluate them
        $x_i = \text{sample\_multivariate\_normal}(\text{mean} = m, \text{ covariance\_matrix} = \sigma^2 C)$
        $f_i = \text{fitness}(x_i)$
    $x_{1\dots\lambda} \leftarrow x_{s(1)\dots s(\lambda)}$ with $s(i) = \text{argsort}(f_{1\dots\lambda}, i)$ // sort solutions
    $m' = m$ // we need later $m - m'$ and $x_i - m'$
    $m \leftarrow \text{update\_m}(x_1, \dots, x_\lambda)$ // move mean to better solutions
    $p_\sigma \leftarrow \text{update\_ps}(p_\sigma, \sigma^{-1} C^{-1/2} (m - m'))$ // update isotropic evolution path
    $p_c \leftarrow \text{update\_pc}(p_c, \sigma^{-1} (m - m'), \|p_\sigma\|)$ // update anisotropic evolution path
    $C \leftarrow \text{update\_C}(C, p_c, (x_1 - m')/\sigma, \dots, (x_\lambda - m')/\sigma)$ // update covariance matrix
    $\sigma \leftarrow \text{update\_sigma}(\sigma, \|p_\sigma\|)$ // update step-size using isotropic path length
return m or x_1

Доменная область

- JEC-QA китайский, законодательство, множ. выбор
  - Knowledge-Driven Questions
  - Case-Analysis Questions
- CaseHOLD — законодательство, анализ прецедентов, множ. выбор
- MLEC-QA китайский, медицина, множ. выбор
  - Clinic
  - Stomatology
  - Public Health
  - Traditional Chinese Medicine
  - Traditional Chinese Medicine Combined with Western Medicine

Бейзлайны

- General LLMs
  - ChatGLM-6B, ChatGLM2-6B, Baichuan-7B/13B-Chat, Baichuan2-7B-Chat/13B-Chat, Qwen-7B-Chat, Chat-GPT
- Legal-specific LLMs
  - LaywerLLaMA (13B), LexiLaw 1 (6B), ChatLaw-13B/33B
- Medical-specific LLMs
  - Taiyi (7B), Zhongjing (13B)
- Retrieval-augmented LLMs
  - BGE-base, M3E-base, GTE-base, piccolo-base

Кто у нас маленькая модель

- https://huggingface.co/bigscience/bloomz-1b7

We recommend using the model to perform tasks expressed in natural language. For example, given the prompt "*Translate to English: Je t'aime.*", the model will most likely answer "*I love you.*". Some prompt ideas from our paper:

- 一个传奇的开端，一个不灭的神话，这不仅仅是一部电影，而是作为一个走进新时代的标签，永远彪炳史册。你认为这句话的立场是赞扬、中立还是批评?
- Suggest at least five related search terms to "Mạng neural nhân tạo".
- Write a fairy tale about a troll saving a princess from a dangerous dragon. The fairy tale is a masterpiece that has achieved praise worldwide and its moral is "Heroes Come in All Shapes and Sizes". Story (in Spanish):
- Explain in a sentence in Telugu what is backpropagation in neural networks.

Что вышло: законодательство

Legal-specific LLMs так себе, и даже хуже после файнтюнинга
BLADE улучшает любую модель.
Accuracy (!)

Model	# Parameters	KD-questions		CA-questions		All	
		Original	+BLADE	Original	+BLADE	Original	+BLADE
Legal Specific LLMs							
LaywerLLaMA	13B	9.76	12.94**(32.6%)	6.05	8.66**(43.1%)	7.45	10.26**(37.7%)
LexiLaw	6B	15.50	19.63**(26.6%)	14.35	18.07**(25.9%)	14.78	18.66**(26.5%)
ChatLaw-13B	13B	10.32	17.32**(67.8%)	5.03	8.08**(60.6%)	7.01	11.55**(64.8%)
ChatLaw-33B	33B	15.66	21.80**(39.2%)	17.01	20.46**(20.3%)	16.50	20.96**(27.0%)
General LLMs							
ChatGLM-6B	6B	17.08	21.19**(24.1%)	16.64	18.62**(11.9%)	16.81	19.58**(16.5%)
ChatGLM2-6B	6B	**27.39**	30.81**(12.5%)	24.09	**26.34****(9.3%)	**25.32**	**28.01****(10.6%)
Qwen-7B-Chat	7B	25.78	**31.26****(21.2%)	**24.52**	25.07*(2.2%)	24.99	27.39**(9.6%)
Baichuan-7B	7B	15.31	21.80**(41.4%)	17.80	21.58**(21.2%)	16.86	21.66**(28.4%)
Baichuan-13B-Chat	13B	17.87	23.06**(14.1%)	19.19	21.71**(13.1%)	18.69	21.21**(13.4%)
Baichuan2-7B-Chat	7B	19.23	24.27**(26.2%)	19.53	21.73**(11.3%)	19.41	22.68**(16.8%)
Baichuan2-13B-Chat	13B	25.78	28.29**(9.73%)	21.80	24.22**(11.1%)	23.29	25.75**(10.5%)
ChatGPT	-	20.53	28.45**(38.6%)	18.70	23.67**(26.6%)	19.38	25.46**(31.3%)

Там еще много таких табличек

Model	Cli		CWM		PH		Sto		TCM	
	Original	+BLADE	Original	+BLADE	Original	+BLADE	Original	+BLADE	Original	+BLADE
Medical Specific LLMs										
Zhongjing_base	15.58	35.74**	19.03	37.52**	16.55	36.98**	14.48	34.86**	17.41	36.65**
Zhongjing_sft	16.00	47.92**	18.50	49.64**	15.85	50.24**	15.76	46.12**	18.88	47.82**
Taiyi	43.42	49.72**	32.71	42.99**	35.11	45.63**	31.53	41.77**	32.83	43.65**
General LLMs										
ChatGLM-6B	30.04	53.42**	30.84	55.06**	30.47	55.66**	27.56	52.24**	32.96	53.64**
ChatGLM2-6B	48.86	60.20**	44.82	57.23**	44.39	59.75**	41.77	57.61**	46.12	55.72**
Qwen-7B-Chat	56.57	59.78*	52.59	58.20**	52.64	62.26**	49.33	57.39**	51.53	56.62**
Baichuan-7B	27.80	54.86**	25.19	56.03**	26.75	58.54**	22.34	50.34**	24.66	52.59**
Baichuan-13B-Chat	42.17	58.98**	45.27	56.59**	42.01	61.54**	38.52	56.42**	41.97	55.66**
Baichuan2-7B-Chat	51.10	59.99**	51.14	58.69**	50.00	62.45**	45.29	57.61**	51.79	56.82**
Baichuan2-13B-Chat	**58.98**	**61.62***	**54.39**	**58.79****	**57.92**	**63.80****	**50.39**	**57.84****	**54.87**	**57.34***
ChatGPT	47.56	58.92**	38.69	57.91**	47.73	63.37**	43.32	57.58**	36.49	56.40**

Некоторые забавные

Table 5: Impact of the number of retrieved documents on JEC-QA. The retrieved corpus is legal_all. Best results are marked bold.

Model	doc_num	KD-questions	CA-questions	All
-	0	27.39	24.09	25.33
M3E_base	1	30.56	24.88	27.02
M3E_base	3	30.71	24.67	26.93
M3E_base	5	30.36	25.29	27.19
M3E_base	7	29.75	24.28	26.33
M3E_base	9	29.63	24.40	26.36
BLADE	-	**30.81**	**26.34**	**28.01**

RAG не вывозит

Query	Zhang has a watch. One day, his friend Li offers to buy it for 30,000 yuan, and Zhang immediately agrees. They sign a contract, with Li agreeing to pay in three installments over two months. After two months, Li has only paid 10,000 yuan. Zhang sues Li for breach of contract. During the trial, Li's parents claim that he has intermittent mental illness and was experiencing an episode when signing the contract, thus unable to be responsible for his actions. They apply to the court to declare Li as a person with limited capacity for civil conduct. How should the court handle this?.....
Retrieved documents	Article 367 of the Criminal Procedure Rules of the People's Procuratorate states that if a criminal suspect is suspected of having mental illness during the review for prosecution, the People's Procuratorate should conduct an assessment according to relevant rules. If the suspect's defender or relatives apply for an assessment due to suspected mental illness, the People's Procuratorate can also conduct an assessment according to the rules, with the costs borne by the applicant……
BLADE	According to the Civil Procedure Law, the determination of whether a party has full civil conduct capacity requires methods like forensic assessment. If a party is declared as having limited capacity for civil conduct during litigation, the contract they signed may be deemed invalid. Before ruling on the contract's validity, it is necessary to determine Li's civil conduct capacity status through a special procedure. The court should suspend the original lawsuit and allow Li's parents to initiate a separate procedure to confirm Li's status as a person with limited civil conduct capacity. Based on the forensic assessment, the court will determine Li's status and make a corresponding judgment……

Figure 4: Comparison of retrieved knowledge with that generated by BLADE.

Все аккуратно сделано

Table 6: Ablation study on JEC-QA under zero-shot setting. The general LLM is ChatGLM2-6B. Best results are marked bold.

Small Model	KD-questions(%)	CA-questions(%)	All(%)
-	27.39	24.09	25.33
BLOOMZ_1b7	26.38	22.40	23.89
+ DP	26.87	23.63	24.85
+ DP & KIT	28.45	24.89	26.23
+ DP & KIT & BPO	**30.81**	**26.34**	**28.01**

Table 7: Impact of sizes on JEC-QA. The general LLM is ChatGLM2-6B. Best results are marked bold.

Small Model	KD-questions(%)	CA-questions(%)	All(%)
-	27.39	24.09	25.33
BLOOMZ_560m	29.05	24.92	26.47
BLOOMZ_1b1	29.80	25.52	27.13
BLOOMZ_1b7	**30.81**	**26.34**	**28.01**

Итого

- Пора учить китайский

5 CONCLUSION

This paper proposes BLADE, a new framework for applying general large language models to new domains. At its core, BLADE employs small language models to assimilate and continually update domain-specific knowledge. The framework solves problems by realizing collaboration between general large language models and a small domain-specific model. It comprises three main stages: Domain-specific Pre-training, Knowledge Instruction Tuning, and Bayesian Prompted Optimization. Domain-specific Pre-training injects domain-specific knowledge into the small model. Knowledge Instruction Tuning activates the instruction-following capacity of the small model. Bayesian Prompted Optimization facilitates better alignment of the small model with the large model. Through ex-