中国人民大學 | 高瓴人工智能学院
RENMIN UNIVERSITY OF CHINA | Gaoling School of Artificial Intelligence

PERPLEXITY-TRAP: PLM-BASED RETRIEVERS OVERRATE LOW PERPLEXITY DOCUMENTS

Haoyu Wang[1,*], Sunhao Dai[1,*], Haiyuan Zhao[1], Liang Pang[2], Xiao Zhang[1]
Gang Wang[3], Zhenhua Dong[3], Jun Xu[1†], Ji-Rong Wen[1]
[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[2]CAS Key Laboratory of AI Safety, Institute of Computing Technology, Beijing, China
[3]Huawei Noah's Ark Lab, Shenzhen, China
{wanghaoyu0924,sunhaodai,haiyuanzhao,zhangx89,junxu}@ruc.edu.cn,
pangliang@ict.ac.cn, {wanggang110, dongzhenhua}@huawei.com

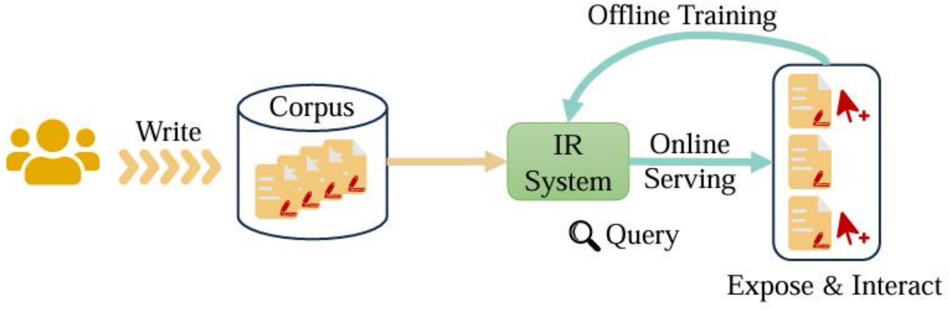
中国人民大学高瓴人工智能学院
Gaoling School of Artificial Intelligence, Renmin University of China
中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES
HUAWEI

Background: IR in the LLM Era

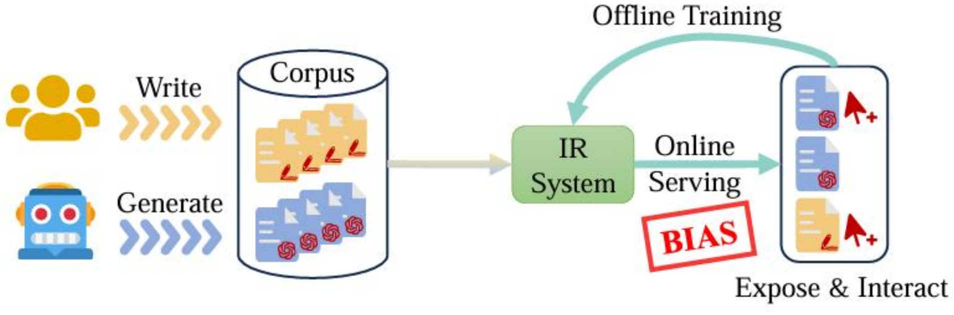
高瓴人工智能学院
Gaoling School of Artificial Intelligence

- AIGC develops by leaps and bound
  - difficult to distinguish
  - Unattributed content sources
- IR systems dominate navigation
  - Control dissemination influence
  - Determine creative income

Previous Research: PLM-based retrieval models exhibit a preference for LLM-generated content, assigning higher relevance scores to AIGC documents.



(a) IR in the Pre-LLM Era



(b) IR in the LLM Era

Source : Dai et al., 2023

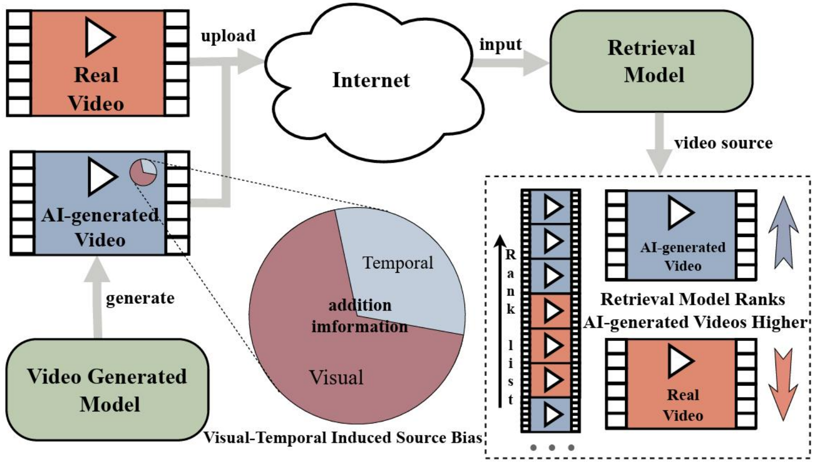
- **Potential Risk: Earn creative incentives through LLM plagiarism**

Dai et al., Neural Retrievers are Biased Towards LLM-Generated Content, KDD 2024.

Source Bias: Common in Many Domains

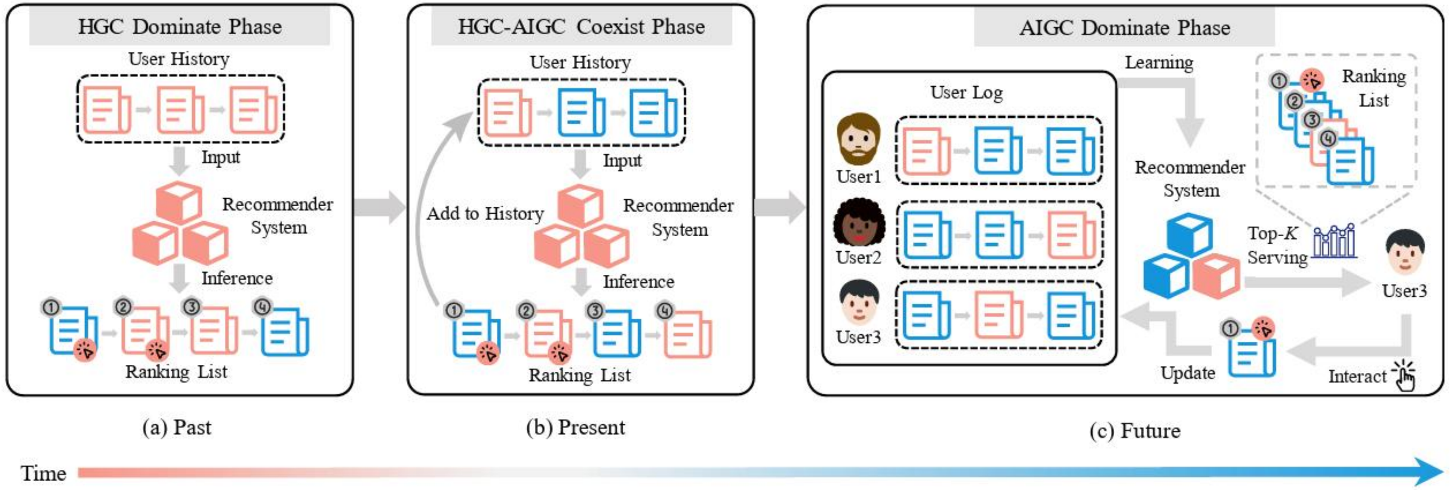
高瓴人工智能学院
Gaoling School of Artificial Intelligence

Video Retrieval



Source: Xu et al., 2024

Feedback Loop in Recommendation



Source: Zhou et al., 2024

Invisible Relevance Bias: Text-Image Retrieval Models Prefer AI-Generated Images, SIGIR 2024
Source Echo Chamber: Exploring the Escalation of Source Bias in User, Data, and Recommender System Feedback Loop.
Generative Ghost: Investigating Ranking Bias Hidden in AI-Generated Videos
Judging llm-as-a-judge with mt-bench and chatbot arena, Neurips 2024

Open Problem: Why Source Bias Occurs

高瓴人工智能学院
Gaoling School of Artificial Intelligence

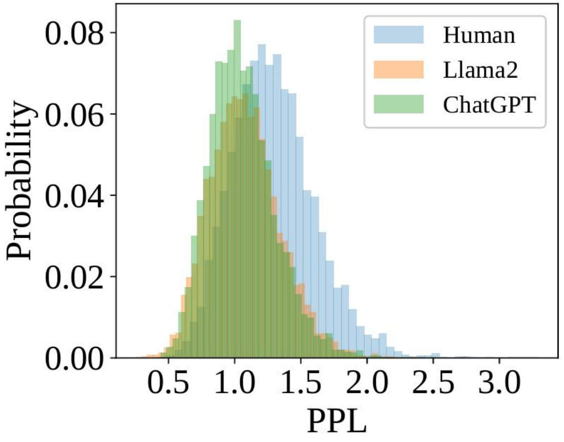
LLM-Generated v.s. Human-Authored

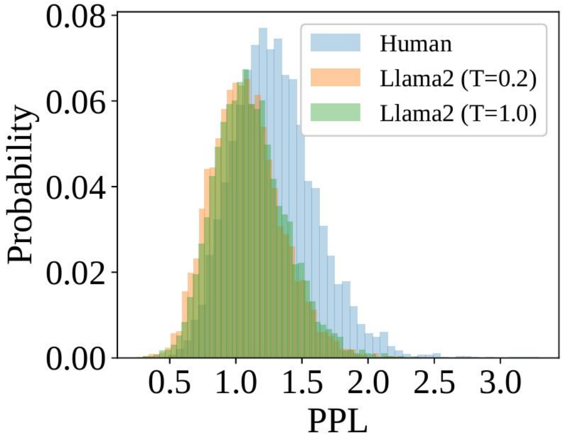
Core Difference: Log Likelihood:

$$\text{Perplexity}(\boldsymbol{x}) \coloneqq \sum_{i=1}^{L} -\log \Pr(\boldsymbol{x}_i | \boldsymbol{x}_{-i})$$

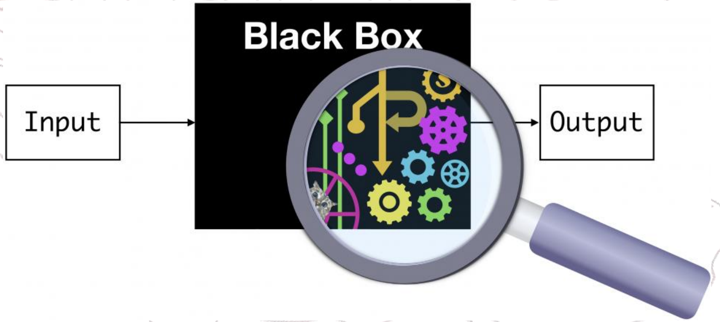
AIGC owns lower perplexity

$$\mathbb{E}_{P_{LLM}(d^G|d^H)}\left[\text{PPL}(d^G, \mathcal{B}) - \text{PPL}(d^H, \mathcal{B})\right] \leq 0.$$





Source: Dai et al., 2023



- ➢ **Whether PPL causally impact Rel. estimation?**
- ➢ **If so, Why causal impact of PPL exists?**
- ➢ **Is it possible to eliminate the effects of PPL?**

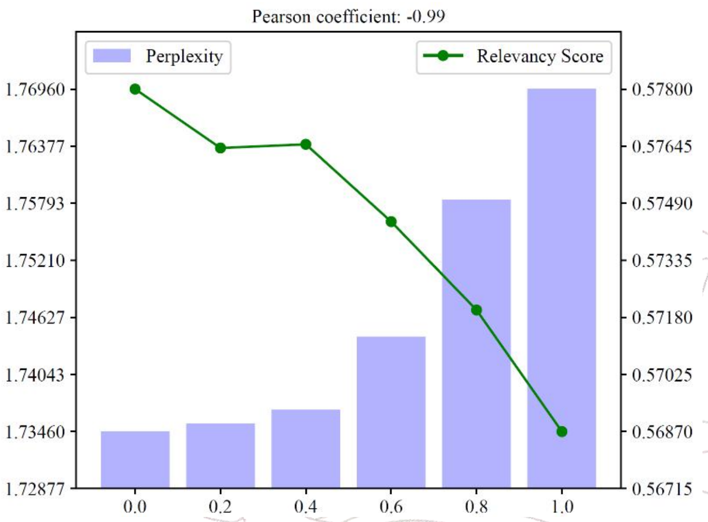
Dai et al., Neural Retrievers are Biased Towards LLM-Generated Content, KDD 2024.

Motivation: Intervention Experiment

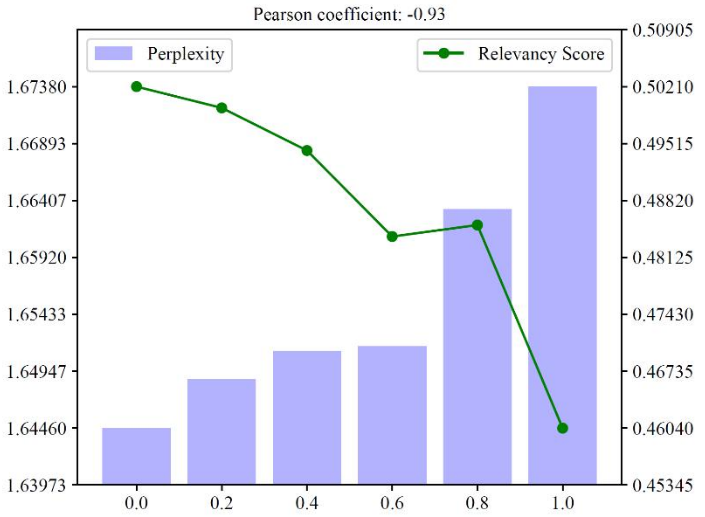
高瓴人工智能学院
Gaoling School of Artificial Intelligence

- Explore the effect of PPL on Rel. ? Control query-document semantics & document source.
- Keep Golden Rel. the same: Only use relevant query-document pairs
- Keep Semantic and Source the same: Only manipulating sampling temperatures.
- **Two variables are highly correlated, there may be causal relationship!**

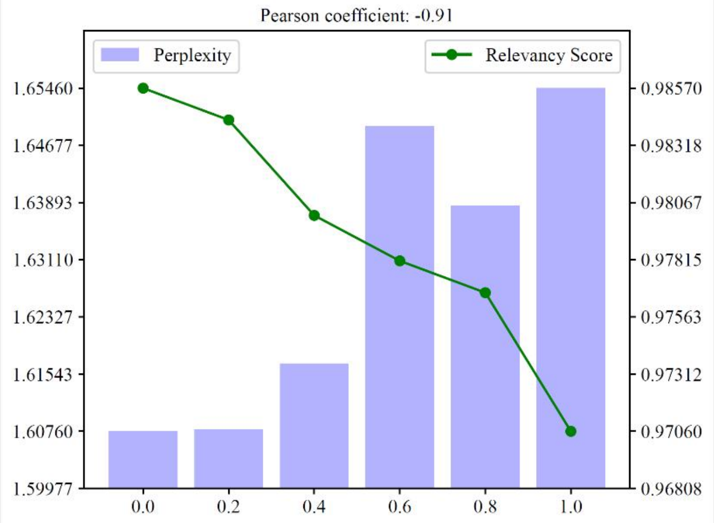
PPL and Rel.(estimated by ANCE) at different sampling temperatures. Pearson coefficients are all lower than -0.9.



a. DL19

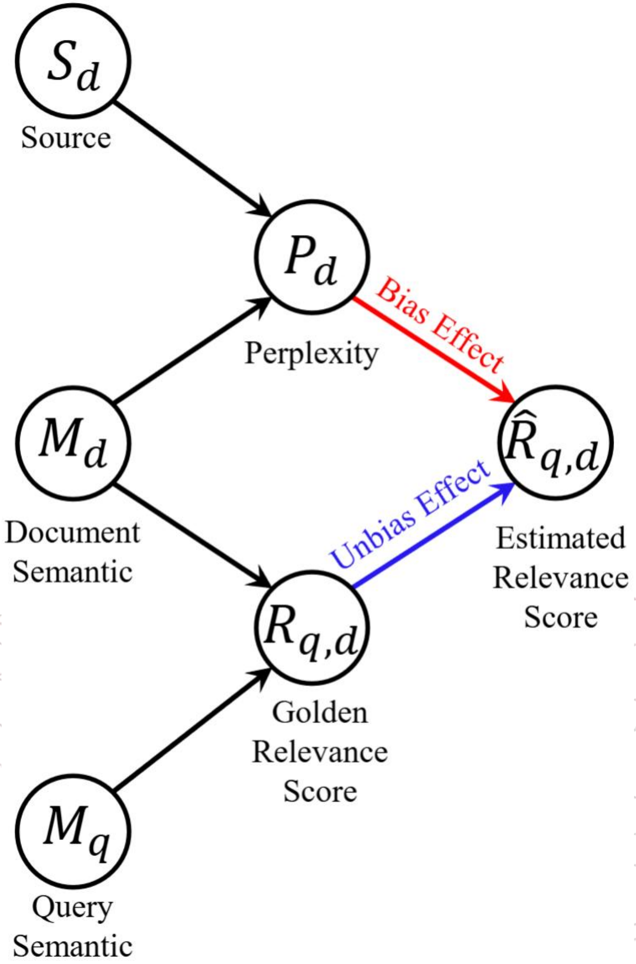


b. NQ



c. TREC-COVID

Hypothesis: Retrieval Causal Graph



$S_d \rightarrow P_d$ AIGC possess lower perplexity than the original document.

$M_d \rightarrow P_d$ Different document semantics lead to different Perplexity.

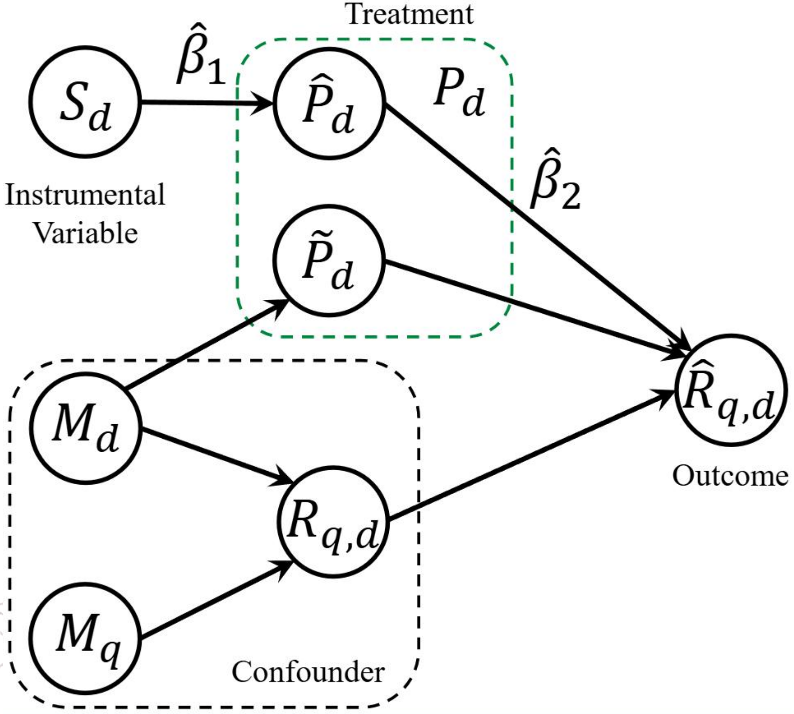
$M_d, M_q \rightarrow R_{q,d}$ Golden Rel. is determined by query-document semantics.

$R_{q,d} \rightarrow \hat{R}_{q,d}$ Retrievers are trained to estimate golden Rel.

$P_d \rightarrow \hat{R}_{q,d}$ Experimental Results indicates this unexpected effect.

Quantify Causal Effect：IV Method

高瓴人工智能学院
Gaoling School of Artificial Intelligence

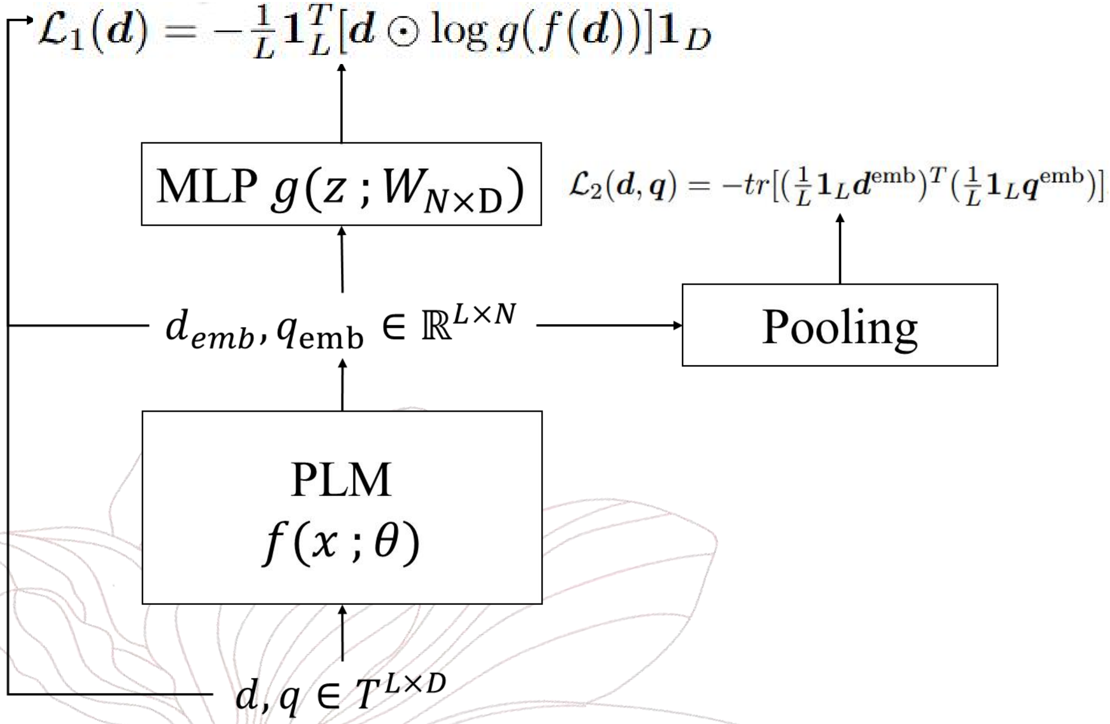


Quantified causal effects (and corresponding *p*-value) for PPL on estimated Rel. Bold indicates *p*-value <0.05. Significant negative causal effects are prevalent across various PLM-based retrievers in different domain datasets.

	BERT	RoBERTa	ANCE
DL19	**-10.42(1e-4)**	**-31.48(2e-12)**	**-0.58(8e-3)**
TREC-COVID	**-1.73(2e-2)**	**2.47(7e-2)**	0.09(0.21)
SCIDOCS	-2.41(6e-2)	**-6.34(2e-3)**	**-0.23(9e-2)**
	TAS-B	**Contriever**	**coCondenser**
DL19	**-1.08(1e-2)**	-0.02(0.33)	**-0.77(3e-2)**
TREC-COVID	**-0.48(5e-3)**	**-0.05(6e-7)**	**-0.33(8e-3)**
SCIDOCS	**-0.39(1e-1)**	-0.02(0.24)	-0.26(0.41)

For PLM-based retrievers, document PPL has a causal effect on estimated Rel. Lower perplexity can lead to higher Rel.

Uncover Mechanism: Theoretical Setting



- **Model Architecture**

Encoder: $f(\boldsymbol{t};\boldsymbol{\theta})\colon \mathcal{T}^{\mathcal{L}\times\mathcal{D}} \mapsto \mathcal{R}^{\mathcal{L}\times\mathcal{N}}$

Decoder: $g(\boldsymbol{z};\boldsymbol{W}) = \sigma(\boldsymbol{z}\boldsymbol{W})$

Simplify: Replace softmax with linear

- **Task Objectives**

Cross Entropy for Masked Language Modeling

$$\mathcal{L}_1(\boldsymbol{d}) = -\frac{1}{L}\mathbf{1}_L^T[\boldsymbol{d} \odot \log g(f(\boldsymbol{d}))]\mathbf{1}_D$$

Mean Pooling & Dot product for document retrieval

$$\mathcal{L}_2(\boldsymbol{d},\boldsymbol{q}) = -tr[(\frac{1}{L}\mathbf{1}_L\boldsymbol{d}^{\mathrm{emb}})^T(\frac{1}{L}\mathbf{1}_L\boldsymbol{q}^{\mathrm{emb}})]$$

MLM & IR Overlap: Aligned Gradients

高瓴人工智能学院
Gaoling School of Artificial Intelligence

Theorem *Given the following three conditions:*

- Representation Collinearity: *the embedding vectors of relevant query-document pairs are collinear after mean pooling*

$$\mathbf{1}_{L\times L} f(\boldsymbol{q}) = \lambda \mathbf{1}_{L\times L} f(\boldsymbol{d}), \lambda > 0.$$

- Semi-Orthogonal Weight Matrix: *decoder weight is semi-orthogonal*

$$\boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{I}_N.$$

- Encoder-decoder Cooperation: *fine-tuning does not disrupt the corresponding function between encoder and decoder*

$$f(\boldsymbol{d}) = g^{-1}(\boldsymbol{d})$$

Then, there exists a matrix: $\boldsymbol{K} = \left[\frac{\lambda k_l}{L(1-k_l)}\right]_{ln} \in \mathcal{R}_+^{L\times N}, k_l = \sum_d^D (\boldsymbol{d}^{\text{emb}}\boldsymbol{W})_{ld}$

$$\frac{\partial \mathcal{L}_2}{\partial \boldsymbol{d}^{\text{emb}}} = \boldsymbol{K} \odot \frac{\partial \mathcal{L}_1}{\partial \boldsymbol{d}^{\text{emb}}}.$$

From Aligned Gradients to Source Bias

$$\frac{\partial \mathcal{L}_2}{\partial \boldsymbol{d}^{\text{emb}}} = \boldsymbol{K} \odot \frac{\partial \mathcal{L}_1}{\partial \boldsymbol{d}^{\text{emb}}}.$$

How to use 1-st order gradients? Taylor Expansion.

Corollary1 *Assume LLM-rewritten documents possess lower perplexity at token level*

$$\mathcal{L}_1^l(\boldsymbol{d}_1) - \mathcal{L}_1^l(\boldsymbol{d}_2) = \frac{\partial \mathcal{L}_1(\boldsymbol{d}_2)}{\partial (\boldsymbol{d}_2^{\text{emb}})_l} \cdot \frac{\partial (\boldsymbol{d}_2^{\text{emb}})_l}{\partial \boldsymbol{d}_2} \cdot \text{vec}(\boldsymbol{d}_1 - \boldsymbol{d}_2) > 0, \;\; l = 1, \ldots, L,$$

According to Theorem 1 and 1st-order approximation of $\mathcal{L}_2(\boldsymbol{d})$

$$\hat{R}_{q,d_1} - \hat{R}_{q,d_2} = -[\mathcal{L}_2(\boldsymbol{d}_1) - \mathcal{L}_2(\boldsymbol{d}_2)] = -\text{rvec}(\boldsymbol{K} \odot \frac{\partial \mathcal{L}_1(\boldsymbol{d}_2^{\text{emb}})}{\partial \boldsymbol{d}_2^{\text{emb}}}) \cdot \frac{\partial \boldsymbol{d}_2^{\text{emb}}}{\partial \boldsymbol{d}_2} \cdot \text{vec}(\boldsymbol{d}_1 - \boldsymbol{d}_2)$$

$$= -\sum_{l=1}^{L} \frac{\lambda k_l}{L(1-k_l)} \frac{\partial \mathcal{L}_1(\boldsymbol{d}_2)}{\partial (\boldsymbol{d}_2^{\text{emb}})_l} \frac{\partial (\boldsymbol{d}_2^{\text{emb}})_l}{\partial \boldsymbol{d}_2} \text{vec}(\boldsymbol{d}_1 - \boldsymbol{d}_2) = -\sum_{l=1}^{L} \frac{\lambda k_l}{L(1-k_l)} \left(\mathcal{L}_1^l(\boldsymbol{d}_1) - \mathcal{L}_1^l(\boldsymbol{d}_2)\right) < 0.$$

Thus, human-written document will receive lower relevance estimation than its LLM-written counterpart.

Deductive Experiments Justification

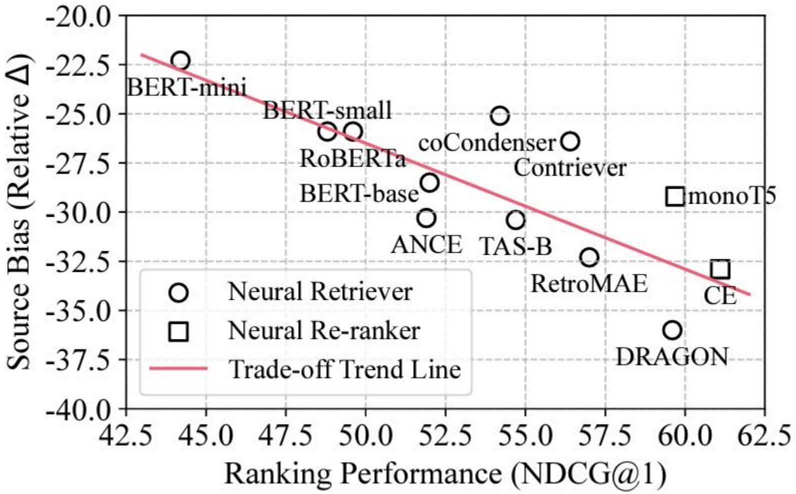
高瓴人工智能学院
Gaoling School of Artificial Intelligence

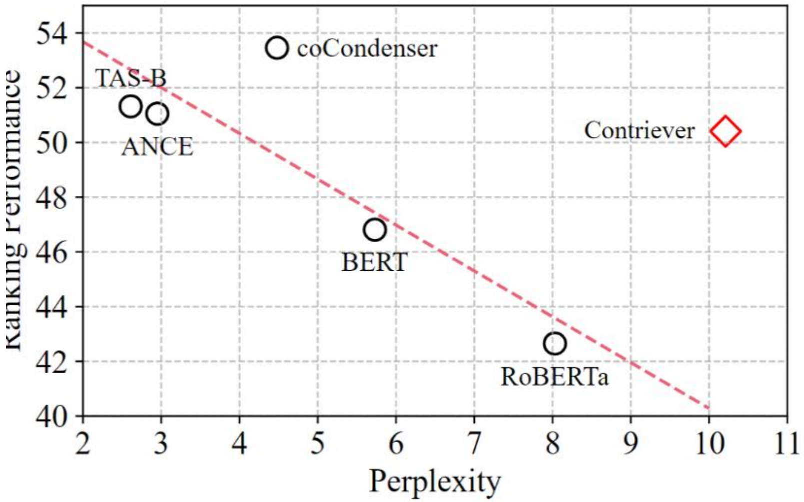
$$\frac{\partial \mathcal{L}_2}{\partial \boldsymbol{d}^{\mathrm{emb}}} = \boldsymbol{K} \odot \frac{\partial \mathcal{L}_1}{\partial \boldsymbol{d}^{\mathrm{emb}}}.$$

Similarly, expansion w.r.t model parameter θ_1 and θ_2

Corollary2 *If retriever $f(t;\ \theta_1)$ possesses more powerful language modeling ability than $f(t;\ \theta_2)$, its ranking performance will be better.*

For PLM-based retrievers, the gradients of MLM and IR loss functions (metrics) possess linear overlap, leading to the biased effect of perplexity on estimated relevance scores.





Source: Dai et al., 2024

Causal Diagnosis and Correction

$$R_{q,d} - \beta_2 P_d = \tilde{R}_{q,d} \perp S_d$$

Algorithm 1: The Proposed CDC: Debiasing with Causal Diagnosis and Correction

Input: training set $\mathcal{D}$, test query set $\mathcal{Q}$, test corpus $\mathcal{C}$, estimation budget M

Output: unbiased estimated relevance scores $\tilde{\mathcal{R}}$

`// Bias Diagnosis`

Initialize the estimation set for estimating biased effect $\mathcal{D}_e \leftarrow \emptyset$

for *training pairs* $(q_i, d_i^{\mathcal{H}}) \in \mathcal{D}$ *and* $|\mathcal{D}_e| < M$ **do**

  Instruct LLM to generate doc $d_i^{\mathcal{G}}$ via rewriting the original human-written doc $d_i^{\mathcal{H}}$

  Predict the estimated relevance scores $\hat{r}_i^{\mathcal{H}}, \hat{r}_i^{\mathcal{G}}$ for pairs $(q_i, d_i^{\mathcal{H}})$ and $(q_i, d_i^{\mathcal{G}})$

  Calculate perplexity $p_i^{\mathcal{H}}, p_i^{\mathcal{G}}$ for doc $d_i^{\mathcal{H}}$ and doc $d_i^{\mathcal{G}}$, respectively

  Updating the estimation set $\mathcal{D}_e \leftarrow \mathcal{D}_e \cup (\hat{r}_i^{\mathcal{H}}, \hat{r}_i^{\mathcal{G}}, p_i^{\mathcal{H}}, p_i^{\mathcal{G}})$

end

Estimate the biased effect coefficient $\hat{\beta}_2$ with 2-stage regression using Eq. (2) on $\mathcal{D}_e$

`// Bias Correction`

for *test query* $q_t \in \mathcal{Q}$ **do**

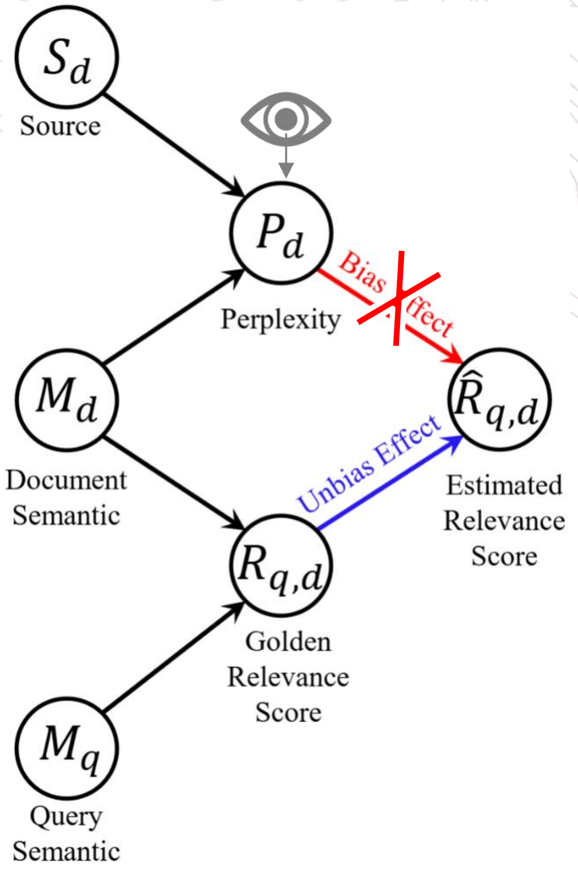
  Predict the estimated relevance scores $\hat{r}_t$ for each pair (q_t, d_t) with $d_t \in \mathcal{C}$

  Calculate document perplexity p_t for each doc $d_t \in \mathcal{C}$

  Debias the original model prediction $\hat{r}_t$ using Eq. (4), add the calibrated score $\tilde{r}_t$ to $\tilde{\mathcal{R}}$

end

return $\tilde{\mathcal{R}}$



Generalization through Data Domains

Training on DL19 and generalize through different data Domains. Eliminate Source bias without significant loss of ranking performance.

Model	DL19 (In-Domain)				TREC-COVID (Out-of-Domain)				SCIDOCS (Out-of-Domain)			
	Performance		Bias		Performance		Bias		Performance		Bias	
	Raw	+CDC	Raw	+CDC	Raw	+CDC	Raw	+CDC	Raw	+CDC	Raw	+CDC
BERT	75.92	77.65	-23.68	5.90	53.72	45.88	-39.58	-18.40	10.80	10.44	-2.85	29.19
Roberta	72.79	71.33	-36.32	4.45	46.31	45.86	-48.14	-10.51	8.85	8.24	-30.90	32.13
ANCE	69.41	67.73	-21.03	34.95	71.01	69.94	-33.59	-1.94	12.73	12.31	-1.57	26.26
TAS-B	74.97	75.63	-49.17	-9.97	63.95	62.84	-73.36	-37.42	15.04	14.15	-1.90	23.48
Contriever	72.61	73.83	-21.93	-5.33	63.17	61.35	-62.26	-31.33	15.45	15.09	-6.96	1.63
coCondenser	75.50	75.36	-18.99	9.60	70.94	71.07	-67.95	-45.39	13.93	13.79	-5.95	1.06

Compare with Constrained Training baseline, achieve comparable results :)

	DL19		TREC-COVID		SCIDOCS	
	Performance	Bias	Performance	Bias	Performance	Bias
Con(0.0001)	62.66	6.25	52.63	46.68	12.76	-8.23
Con(0.0005)	62.69	118.83	51.35	39.10	12.45	26.91
Con(0.001)	62.66	127.25	45.43	85.54	12.41	56.31
Con(0.005)	61.17	175.47	54.00	163.41	10.70	118.87
Con(0.01)	57.62	175.86	39.69	179.84	11.30	111.51
CDC	67.73	34.95	67.94	-1.94	12.31	26.26

Separate biased effect of perplexity is effective and efficient for source bias.

Generalization through LLM Rewriters

高瓴人工智能学院
Gaoling School of Artificial Intelligence

CDC show good generalization across different LLMs.

Model	Llama-2 (In-Domain)				GPT-4 (Out-of-Domain)				GPT-3.5 (Out-of-Domain)				Mistral (Out-of-Domain)			
	Performance		Bias		Performance		Bias		Performance		Bias		Performance		Bias	
	Raw	+CDC	Raw	+CDC	Raw	+CDC	Raw	+CDC	Raw	+CDC	Raw	+CDC	Raw	+CDC	Raw	+CDC
BERT	35.67	35.08	-12.37	6.75	36.47	35.75	-3.69	6.04	35.97	35.27	-5.03	18.08	35.13	35.08	0.73	13.07
RoBERTa	38.09	36.76	-29.54	-0.88	38.53	37.70	-11.98	4.52	39.17	38.00	-35.39	14.09	38.29	37.28	-17.95	16.78
ANCE	42.13	42.13	-8.81	4.59	42.67	42.99	-5.53	3.28	42.76	42.96	-13.59	6.09	42.62	42.71	-8.59	1.82
TAS-B	52.95	53.94	-15.04	-7.96	52.12	52.44	-4.94	-0.05	52.83	52.90	-5.65	5.57	52.18	52.69	-8.71	-2.00
Contriever	55.19	55.37	-2.87	1.07	55.78	55.70	-5.32	-4.44	56.11	56.17	-7.43	-2.81	56.13	56.28	-4.13	-2.39
coCondenser	49.53	49.40	-12.98	-9.26	48.57	48.91	5.04	6.04	48.59	48.81	-1.00	5.30	49.57	49.92	-5.90	-0.76

- Much fewer training data while good debiasing performance.
- Well Generalization ability to adapt real scenario.
- Document perplexity can be computed and indexed offline without increasing online latency.
- Controllable trade-off between retrieval accuracy and unbiasedness by adjusting $\widehat{\beta}_2$.

Thanks!

Contact if you have any question :)
wanghaoyu0924@ruc.edu.cn