

KD字典不一致

分类	方法	总结
概率分布对齐	ULD	用 最优传输 (Optimal Transport) 的 Wasserstein 距离直接比较教师与学生的 输出分布 ，即直接让学生去学校教师的概率分布
	DSKD	学教师的 概率分布 ，还学教师隐藏层里的 语义关系 ，从教师与学生的 隐藏层 提取语义嵌入
多层次对齐 (概率、结构、语义)	MultiLevelOT	多层次 最优传输 在 token 层 (概率分布差异和比例差异) 和句子层 (语义) 同时对齐， Sinkhorn Distance 加快计算
	CMD	用 Dynamic Time Warping 自动对齐局部 (加权重要的 token)，再构建上下文相关的动态语义映射矩阵，映射函数基于相邻上下文token的语义表示 (全局语义)
	EMO	注意力结构相似度 (结构相似)，最后一层隐藏状态进行 语义对齐 (语义相似)， 最优传输 计算从教师分布到学生分布的最小代价
	CoT2Align	学生学老师思考的过程 (Chain of Thought)，把推理链条也蒸馏下来
文本对齐	VocAgnoLM	直接对齐原文字符位置，通过 字符偏移 找到教师和学生的对应 token
	ALM	把不同 tokenizer 切出来的文本分成语义相同的块，对齐在 相同意思的文本片段 上的概率

概率分布对齐

[Towards Cross-Tokenizer Distillation: the Universal Logit Distillation Loss for LLMs](#)

[Nicolas Boizard](#), [Kevin El Haddad](#), [Céline Hudelot](#), [Pierre Colombo](#)

方法 ULD

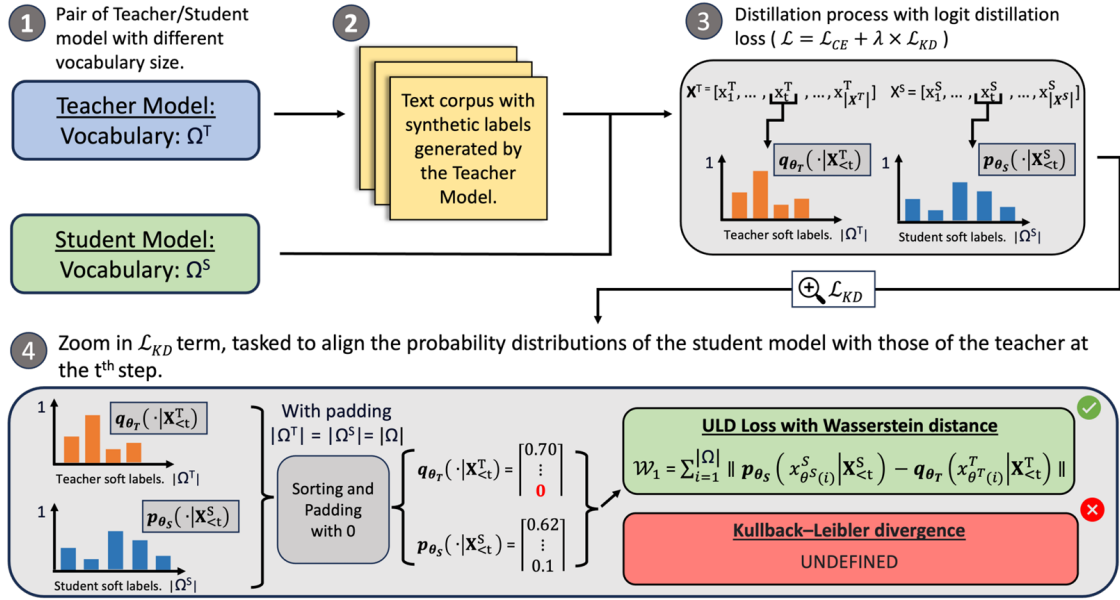


Figure 2: **Distillation using ULD loss.** In block 4, the KL divergence cannot be defined as the two distributions do not have the same support, breaking the absolute continuity of the quotient in the KL logarithmic term. To alleviate this we rely on the ULD loss which leverages a closed form of the Wasserstein distance.

提出基于 **最优传输 (Optimal Transport, OT) 理论** 的通用蒸馏损失函数，通过计算 **Wasserstein 距离** 来衡量教师与学生输出概率分布之间的差异，从而不再依赖 token 对齐。

(1) 传统KD公式 $L = L_{CE} + \lambda L_{KD}$

(2) ULD损失 $L_{ULD} = \sum_{t=1}^{|x|} CE(t) + \lambda W_1[p_{\theta_S}(\cdot | x_S^{<t}), q_{\theta_T}(\cdot | x_T^{<t})]$

其中：

- W_1 Wasserstein 距离；
- 通过排序后的概率差计算闭式形式，复杂度降至 $O(n \log n)$ 。

(3) 计算优化

为避免 $(O(n^3))$ 复杂度的OT求解，论文提出：

- **Uniform Support Length**：填充词表使维度一致；
- **Uniform Cost Matrix**：假设所有 token 间传输成本相同；
- **闭式快速解**： $W_1(p, q) = \sum_i |p_{\sigma_S(i)} - q_{\sigma_T(i)}|$ (σ 表示按概率从高到低排序)

这样ULD Loss 可以在标准 GPU 上高效计算。

效果

Table 2: **Overall performance** of Teacher/Student pair models trained with ULD Loss and teacher-generated text (Raw Text) across tasks with their main metrics. Evaluations are performed over respective test splits.

Teacher	Model	Method	SQUAD (F1)	QED (F1)	FairytaleQA (BERTScore)	PubMedQA (BERTScore)	DIALOGSum (Rouge-Lsum)
Teacher	LLama	-	81.30	57.72	41.59	30.62	23.90
	Mistral	-	76.31	53.01	36.01	30.93	34.71
LLama	OPT-350m	Raw Text	70.78	48.64	33.78	27.99	20.58
		ULD Loss	72.97	49.06	33.03	30.01	20.11
	Pythia-410m	Raw Text	71.39	47.04	33.02	29.86	20.94
		ULD Loss	74.14	49.15	34.83	29.89	22.19
	Bloomz-560m	Raw Text	73.54	50.99	36.70	29.14	20.01
		ULD Loss	75.90	55.33	37.86	30.01	22.67
	OPT-350m	Raw Text	71.64	50.13	30.09	27.91	31.44
		ULD Loss	73.35	50.88	30.44	30.30	32.17
Mistral	Pythia-410m	Raw Text	71.50	47.07	31.44	28.25	31.64
		ULD Loss	73.64	50.38	31.79	29.55	33.10
	Bloomz-560m	Raw Text	73.34	52.15	32.64	28.87	31.95
		ULD Loss	76.00	55.79	33.93	30.60	32.58
Average	-	Raw Text	72.03	49.34	32.94	28.67	26.09
	-	ULD Loss	74.33	51.77	33.65	30.06	27.14

Dual-Space Knowledge Distillation for Large Language Models

EMNLP 2024, Songming Zhang, Xue Zhang, Zengkui Sun, Yufeng Chen, Jinan Xu

方法 DSKD

它在 **两个空间 (dual-space)** 中实现教师与学生的知识对齐：

1. 概率空间蒸馏 (Probability-space Distillation) ,**基于相似度分布的跨词表对齐机制**：

- 对教师输出进行 softmax 得到概率分布 P_T ;
- 对学生输出进行 softmax 得到 P_S ;
- 将二者通过**分布间相似性投影**映射到一个共享的“分布嵌入空间”;
- 再计算它们之间的 **KL散度** 或 **交叉熵损失**。

“不同词表但相似语义”下建立柔性对齐。

2. 语义空间蒸馏 (Embedding-space Distillation) 为了让学生不仅模仿输出概率，还能学习教师的上下文语义结构,DSKD 在**词级别嵌入空间**中进行额外的知识对齐。

- 从教师与学生的**隐藏层**提取语义嵌入;
- 对应位置的嵌入向量通过 **cosine similarity 损失** 对齐;
- 并引入**动态权重机制**：在重要语义位置（如实体、关键词）加大对齐权重。

这让学生在语义层面接近教师

3. Dual-Space 整体损失函数 $\mathcal{L} * DSKD = \lambda_1 \mathcal{L} * prob + \lambda_2 \mathcal{L}_{emb}$ λ_1, λ_2 为权重超参数。

效果

Methods	Dolly	SelfInst	VicunaEval	S-NI	UnNI	Avg.
SFT	22.94 \pm 0.28	10.11 \pm 0.36	15.17 \pm 0.63	16.21 \pm 0.19	18.68 \pm 0.09	16.62
GPT2-1.5B \rightarrow GPT2-120M (Same Vocabulary)						
Teacher	27.19 \pm 0.23	14.64 \pm 0.64	16.30 \pm 0.37	27.55 \pm 0.30	31.42 \pm 0.11	23.42
SeqKD	23.68 \pm 0.25	10.03 \pm 0.23	14.41 \pm 0.46	16.36 \pm 0.18	18.48 \pm 0.11	16.59
KL	24.54 \pm 0.48	10.43 \pm 0.24	15.66 \pm 0.42	17.24 \pm 0.27	20.28 \pm 0.18	17.63
w/ DSKD (ours)	24.70 \pm 0.24	10.65 \pm 0.30	15.67 \pm 0.30	19.51 \pm 0.21	22.94 \pm 0.07	18.69 (+1.06 \uparrow)
RKL	24.38 \pm 0.55	10.73 \pm 0.61	15.71 \pm 0.39	17.31 \pm 0.11	20.96 \pm 0.12	17.82
w/ DSKD (ours)	24.61 \pm 0.59	11.01 \pm 0.45	14.98 \pm 0.48	19.32 \pm 0.28	22.27 \pm 0.13	18.44 (+0.62 \uparrow)
JS	23.86 \pm 0.14	10.20 \pm 0.40	15.50 \pm 0.23	16.20 \pm 0.23	19.17 \pm 0.06	16.98
w/ DSKD (ours)	24.61 \pm 0.27	11.41 \pm 0.35	15.40 \pm 0.28	18.94 \pm 0.20	21.48 \pm 0.17	18.37 (+1.39 \uparrow)
SKL (Ko et al., 2024)	24.03 \pm 0.23	10.66 \pm 0.51	14.70 \pm 0.37	17.99 \pm 0.15	21.18 \pm 0.16	17.71
w/ DSKD (ours)	25.24 \pm 0.28	10.50 \pm 0.13	15.76 \pm 0.43	18.34 \pm 0.44	20.87 \pm 0.11	18.14 (+0.43 \uparrow)
SRKL (Ko et al., 2024)	24.48 \pm 0.19	10.35 \pm 0.38	14.88 \pm 0.24	16.53 \pm 0.23	19.68 \pm 0.05	17.19
w/ DSKD (ours)	25.23 \pm 0.25	11.19 \pm 0.22	15.91 \pm 0.45	17.92 \pm 0.16	21.20 \pm 0.12	18.29 (+1.10 \uparrow)
AKL (Wu et al., 2024)	24.75 \pm 0.60	10.46 \pm 0.24	15.37 \pm 0.41	17.48 \pm 0.17	20.11 \pm 0.05	17.63
w/ DSKD (ours)	25.13 \pm 0.14	10.63 \pm 0.43	16.18 \pm 0.35	18.58 \pm 0.48	21.45 \pm 0.16	18.39 (+0.76 \uparrow)
Qwen1.5-1.8B \rightarrow GPT2-120M (Different Vocabularies)						
Teacher	27.42 \pm 0.33	19.42 \pm 0.11	19.31 \pm 0.21	34.87 \pm 0.30	36.00 \pm 0.10	27.40
SeqKD	23.40 \pm 0.21	9.36 \pm 0.38	15.37 \pm 0.35	15.16 \pm 0.17	17.34 \pm 0.11	16.13
MinED (Wan et al., 2024)	24.41 \pm 0.61	10.60 \pm 0.39	15.86 \pm 0.42	16.76 \pm 0.28	19.68 \pm 0.12	17.46
ULD (Boizard et al., 2024)	23.77 \pm 0.41	9.67 \pm 0.50	14.99 \pm 0.55	17.60 \pm 0.21	19.49 \pm 0.12	17.11
DSKD-CMA-SRKL (ours)	25.23 \pm 0.17	10.99 \pm 0.26	15.56 \pm 0.41	17.76 \pm 0.23	20.54 \pm 0.07	18.02

Table 1: Rouge-L scores (%) on several benchmarks with GPT2-120M as the student. We list the mean values and the standard deviations among 5 random seeds. The average scores (Avg.) on all benchmarks are also listed. “w/ DSKD” denotes our DSKD using the corresponding distance function as $\mathcal{D}(\cdot||\cdot)$ in Eqn. (6). And “DSKD-CMA-SRKL” denotes our DSKD framework equipped with cross-model attention with SRKL as $\mathcal{D}(\cdot||\cdot)$ in Eqn. (6).

Methods	Dolly	SelfInst	VicunaEval	S-NI	UnNI	Avg.
SFT	23.20 \pm 0.13	14.88 \pm 0.54	16.42 \pm 0.35	27.79 \pm 0.27	26.12 \pm 0.11	21.68
LLaMA2-7B \rightarrow TinyLLaMA-1.1B (Same Vocabulary)						
Teacher	28.32 \pm 0.46	20.95 \pm 0.69	18.76 \pm 0.35	32.05 \pm 0.28	32.41 \pm 0.12	26.50
SeqKD	23.21 \pm 0.22	16.46 \pm 0.72	16.58 \pm 0.38	26.33 \pm 0.26	27.69 \pm 0.10	22.05
KL	25.46 \pm 0.63	17.21 \pm 0.25	16.43 \pm 0.53	29.27 \pm 0.29	29.28 \pm 0.09	23.53
w/ DSKD (ours)	26.31 \pm 0.26	18.27 \pm 0.56	18.04 \pm 0.37	31.43 \pm 0.26	31.20 \pm 0.09	25.05 (+1.52 \uparrow)
RKL	24.49 \pm 0.41	17.14 \pm 0.61	16.87 \pm 0.26	29.50 \pm 0.28	29.36 \pm 0.08	23.47
w/ DSKD (ours)	26.93 \pm 0.34	18.14 \pm 0.54	18.81 \pm 0.39	31.79 \pm 0.31	32.49 \pm 0.11	25.63 (+2.17 \uparrow)
JS	24.03 \pm 0.31	15.75 \pm 0.51	16.64 \pm 0.30	28.08 \pm 0.10	28.68 \pm 0.08	22.62
w/ DSKD (ours)	24.79 \pm 0.42	17.10 \pm 0.47	16.78 \pm 0.20	29.06 \pm 0.18	29.47 \pm 0.22	23.44 (+0.82 \uparrow)
SKL (Ko et al., 2024)	24.14 \pm 0.53	15.98 \pm 0.72	16.89 \pm 0.22	29.30 \pm 0.18	28.71 \pm 0.12	23.01
w/ DSKD (ours)	25.88 \pm 0.22	17.59 \pm 0.56	17.17 \pm 0.34	29.52 \pm 0.33	30.69 \pm 0.16	24.17 (+1.16 \uparrow)
SRKL (Ko et al., 2024)	24.28 \pm 0.58	16.91 \pm 0.67	16.88 \pm 0.20	29.55 \pm 0.19	28.64 \pm 0.21	23.25
w/ DSKD (ours)	25.44 \pm 0.22	17.34 \pm 0.69	17.19 \pm 0.34	30.29 \pm 0.29	31.23 \pm 0.13	24.30 (+1.05 \uparrow)
AKL (Wu et al., 2024)	24.80 \pm 0.70	16.79 \pm 1.09	16.80 \pm 0.44	29.29 \pm 0.35	28.81 \pm 0.09	23.30
w/ DSKD (ours)	26.33 \pm 0.45	20.17 \pm 0.46	17.43 \pm 0.48	34.93 \pm 0.39	34.40 \pm 0.20	26.65 (+3.35 \uparrow)
Mistral-7B \rightarrow TinyLLaMA-1.1B (Different Vocabularies)						
Teacher	31.56 \pm 0.19	25.10 \pm 0.36	20.50 \pm 0.32	36.07 \pm 0.24	36.27 \pm 0.15	29.90
SeqKD	23.56 \pm 0.39	15.87 \pm 0.54	15.99 \pm 0.55	25.50 \pm 0.37	26.64 \pm 0.09	21.51
MinED (Wan et al., 2024)	20.96 \pm 0.51	14.49 \pm 0.35	15.98 \pm 0.45	27.21 \pm 0.13	26.47 \pm 0.11	21.77
ULD (Boizard et al., 2024)	22.80 \pm 0.28	15.93 \pm 0.74	16.43 \pm 0.60	26.94 \pm 0.28	24.83 \pm 0.13	20.64
DSKD-CMA-AKL (ours)	26.45 \pm 0.56	19.57 \pm 0.69	17.95 \pm 0.55	35.99 \pm 0.19	35.00 \pm 0.16	26.99

Table 2: Rouge-L scores (%) on several benchmarks with TinyLLaMA-1.1B as the student. We list the mean values and the standard deviations among 5 random seeds. “w/ DSKD” denotes our DSKD using the corresponding distance function as $\mathcal{D}(\cdot||\cdot)$ in Eqn. (6). And “DSKD-CMA-AKL” denotes our DSKD framework equipped with

多层次对齐

Multi-Level Optimal Transport for Universal Cross-Tokenizer Knowledge Distillation on Language Models

AAAI 2025 (Oral), Xiao Cui, Mo Zhu, Yulei Qin, Liang Xie, Wengang Zhou, Houqiang Li

方法 MultiLevelOT

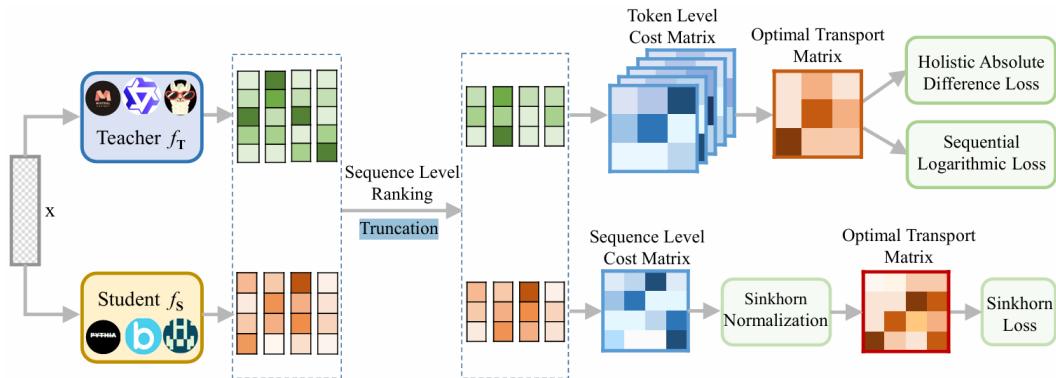


Figure 2: Illustration of our pipeline. MultiLevelOT computes sequence-aware token-level and sequence-level optimal transport distances between the output logits of the teacher and student models. This approach effectively transfers local and global information within the logits distribution, accommodating vocabulary differences and enabling cross-tokenizer distillation.

MultiLevelOT 两个层次的对齐：

1. **Token-level OT (词元层面)**：对每个句子的所有词一起计算分布差异，而不是逐个 token。它使用两种互补的代价矩阵：
 - **绝对差代价矩阵 (L1距离, 衡量直接差异)**：直接比较教师和学生输出的差异
 - **对数差代价矩阵 (Log-diff, 捕捉比例关系)**：捕捉相对差异，对不同量级的输出更加敏感
2. **Sequence-level OT (句子层面)**：在整个序列范围上，再用最优传输度量整体语义差异，能自动应对不同token切分方式造成的错位问题。

同时使用 **Sinkhorn Distance** (一种高效的Wasserstein距离近似) 加快计算，在保持语义结构信息的同时显著降低计算量。

效果

Model	Method	QED (F1)	FairytaleQA (Rouge-LSum)	DIALOGSum (Rouge-LSum)
LLaMA2-7B	Few-Shot	61.68	50.90	37.75
OPT-350M	Origin	12.46	11.16	14.02
	SFT	55.71	46.04	35.59
	SeqKD	49.61	39.19	30.71
	MinED	56.03	46.11	35.82
	ULD	56.76	45.82	36.05
	Ours	58.97	46.96	37.61
Pythia-410M	Origin	22.87	15.14	4.41
	SFT	59.03	47.23	36.06
	SeqKD	51.12	39.78	31.57
	MinED	59.21	47.31	35.97
	ULD	59.71	47.81	36.07
	Ours	61.79	49.10	37.45
Bloomz-560M	Origin	47.67	43.47	11.82
	SFT	60.48	49.07	36.52
	SeqKD	52.33	45.68	31.83
	MinED	60.52	49.10	36.39
	ULD	61.22	49.87	36.40
	Ours	62.58	50.94	37.68
Average	Origin	27.67	23.25	10.08
	SFT	58.41	47.45	36.05
	SeqKD	50.99	41.55	31.37
	MinED	58.58	47.47	36.06
	ULD	59.30	47.83	36.17
	Ours	60.99	49.00	37.58

Table 1: Performance of the students in labeled distillation. Both the teacher and ground-truth provide supervision.

Model	Method	QED (F1)	FairytaleQA (Rouge-LSum)	DIALOGSum (Rouge-LSum)
LLaMA2-7B	Few-Shot	61.68	50.90	37.75
OPT-350M	Origin	12.46	11.16	14.02
	Raw Text	49.61	39.19	30.71
	ULD	50.71	39.86	32.03
	Ours	51.96	40.68	36.88
Pythia-410M	Origin	22.87	15.14	4.41
	Raw Text	51.12	39.78	31.57
	ULD	52.09	40.69	34.15
	Ours	53.56	41.28	36.52
Bloomz-560M	Origin	47.67	43.47	11.82
	Raw Text	52.33	45.68	31.83
	ULD	53.02	46.72	34.21
	Ours	54.15	47.88	37.10
Average	Origin	27.67	23.25	10.08
	Raw Text	50.99	41.55	31.37
	ULD	51.94	42.42	33.46
	Ours	53.22	43.28	36.83

Table 2: Performance of the students in unlabeled distillation. The ground-truth is unavailable for supervision.

Enhancing Cross-Tokenizer Knowledge Distillation with Contextual Dynamical Mapping

方法 CMD

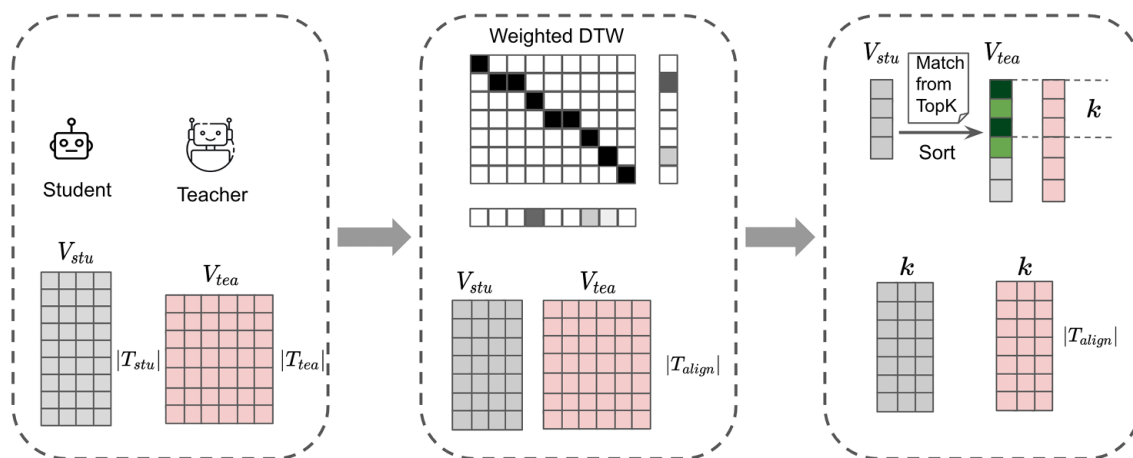


Figure 4: The architecture of CDM consists of two key components: an entropy-weighted Dynamic Time Warpin (DTW) sequence alignment algorithm and a dynamic Top-K vocabulary mapping algorithm. Following the mappin procedure, the output representations from both the teacher and student models are aligned to ensure consistency i both dimensional structure and semantic space.

1. 动态时间规整 (Dynamic Time Warping, DTW) 使用 DTW 来匹配教师和学生 token 在时间维度 (token 序列) 上的语义变化轨迹, 自动对齐不同分词方式下语义相近的部分;
- 引入 熵加权机制 (Entropy-weighted DTW) :
 - 对教师输出分布的不确定性 (entropy) 加权,
 - 使得语义更清晰的 token 在对齐中占更大权重;
 - 提升对齐稳定性和鲁棒性。
2. 上下文动态映射 (Contextual Dynamical Mapping, CDM) 进一步构建 上下文相关的动态语义映射矩阵
 - 该矩阵在训练中自动学习每个教师 token 对应学生 token 的语义相关性;
 - 映射函数基于相邻上下文 token 的语义表示, 动态调整;
 - 形式上为: $M_{i,j} = f(h_i^T, h_j^S, C_i, C_j)$
其中 C_i, C_j 表示局部上下文表示, h_i^T, h_j^S 分别为教师与学生的隐藏状态。

CDM 让学生不需要知道教师具体的 token, 而是学习教师在上下文语义层面上的模式。

3. 语义一致性蒸馏损失 (Contextual Distillation Loss) $L_{CDM} = L_{CE} + \lambda_1 L_{DTW} + \lambda_2 L_{CM}$

效果

Type	Setting	Dolly	Self-Inst	Vicuna	S-NI	UnNI	#Avg IF	HumanEval+	MBPP+	GSM-8K
Student Model: Gemma-2-2B										
SFT	Gemma-2-2B	25.12	14.94	16.89	25.29	30.07	22.46	21.34	21.34	29.95
	Gemma-2-9B	26.72	18.01	18.85	27.74	34.83	25.23	24.39	27.51	45.34
	Llama-3-8B	27.01	21.90	17.00	30.66	35.23	26.36	34.76	50.26	44.20
Same Tokenizer KD (teacher: Gemma-2-9B)	FKL	26.51	14.30	18.64	27.61	32.06	23.82	18.90	23.00	34.80
	RKL	25.26	13.80	18.64	23.70	29.79	22.24	18.90	21.42	27.37
Cross Tokenizer KD (teacher: Llama-3-8B)	MinED	25.83	16.16	16.40	25.99	28.60	22.60	20.12	22.22	28.43
	ULD	26.11	14.58	17.25	27.69	30.53	23.23	20.40	17.70	26.38
	CDM	26.13	14.89	18.33	26.40	32.00	23.55	23.78	21.69	30.40
Student Model: OPT-1.3B										
SFT	OPT-1.3B	25.48	14.26	14.81	25.88	31.93	22.47	-	-	-
	OPT-6.7B	28.40	15.71	15.82	26.87	33.56	24.07	-	-	-
	Llama-3-8B	27.01	21.90	17.00	30.66	35.23	26.36	-	-	-
Same Tokenizer KD (OPT-6.7B)	FKL	25.36	15.24	16.16	26.47	31.38	22.92	-	-	-
	RKL	25.03	13.24	15.42	23.86	31.27	21.77	-	-	-
Cross Tokenizer KD (teacher: Llama-3-8B)	MinED	25.21	12.60	15.60	24.51	30.52	21.69	-	-	-
	ULD	25.45	13.69	15.88	25.82	30.07	22.18	-	-	-
	CDM	26.15	14.39	15.77	26.33	32.33	23.00	-	-	-
Student Model: Qwen2-0.5B										
SFT	Qwen2-0.5B	24.66	15.17	15.22	30.31	35.00	24.07	15.85	22.22	27.22
	Qwen2-7B	29.07	22.69	21.42	37.31	41.04	30.31	39.02	39.42	59.14
	Phi3-mini-3.8B	29.19	25.39	21.81	37.97	41.07	31.09	51.83	48.68	64.67
Same Tokenizer KD (Qwen2-7B)	FKL	27.41	19.68	19.24	32.67	37.46	27.29	17.07	23.38	27.67
	RKL	26.15	16.15	16.62	30.32	37.53	25.35	20.73	22.75	26.38
Cross Tokenizer KD (teacher: Phi3-mini-3.8B)	MinED	25.55	16.26	15.37	30.76	35.69	24.72	17.10	22.20	24.41
	ULD	26.43	16.15	15.34	30.63	36.07	24.93	17.07	22.49	26.38
	CDM	25.45	16.55	16.38	30.66	36.47	25.10	18.90	23.81	28.13

Table 3: Main results of comparing CDM and the baseline models, where“#AVG IF” means the average score of the instruction-following tasks). The **blod** text means the best performance in comparable cross-tokenizer distillation settings. The table consists of three sections, each labeled with the student models in distillation experiments.

好的，下面是对论文 **Universal Cross-Tokenizer Distillation via Approximate Likelihood Matching** (arXiv: 2503.20083) 的方法与效果的详细说明。

[EMO: Embedding Model Distillation via Intra-Model Relation and Optimal Transport Alignments - ACL Anthology](#)

EMNLP 2025, [Minh-Phuc Truong](#), [Hai An Vu](#), [Tu Vu](#), [Nguyen Thi Ngoc Diep](#), [Linh Ngo Van](#), [Thien Huu Nguyen](#), [Trung Le](#)

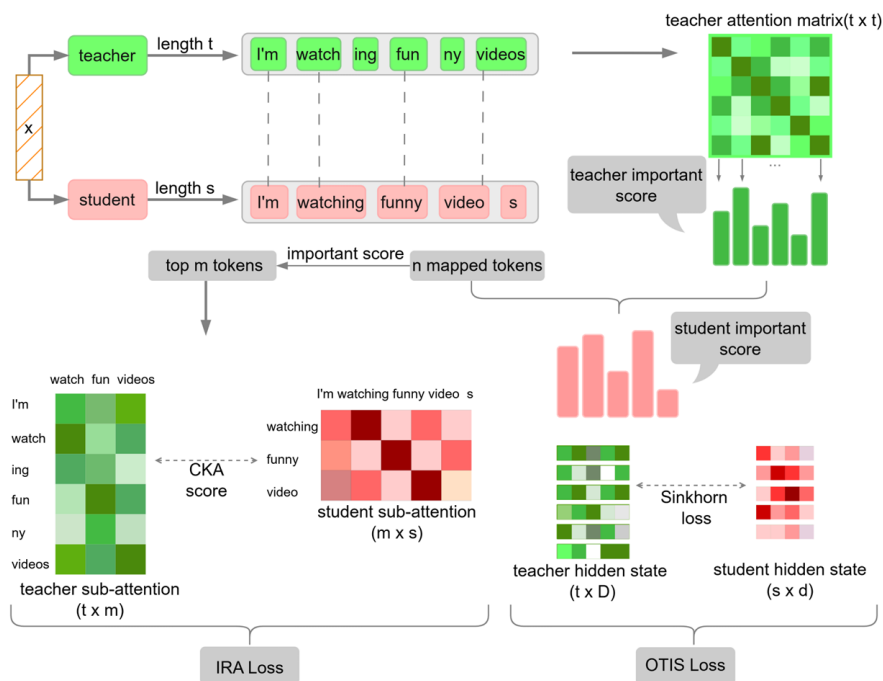


Figure 1: Overall workflow of our EMO framework. We perform Intra-Model Relational Distillation (IRA) using MinED for token mapping, followed by CKA on attention matrices and employ Optimal Transport with Importance-Scored Mass Assignment (OTIS) for cross-model representation alignment.

方法 EMO

让学生模型在**语义与结构层面**对齐教师模型。

1. 使用 Minimum Edit Distance (MinED) 找到教师与学生序列中“**最相似**”的 **token 对**，建立基础的 **1-1 token 映射**。
2. 内部结构关系蒸馏 (Intra-Model Relation Alignment, IRA)
 1. 对已匹配的 token，利用教师模型的*注意力矩阵) 计算哪些 token 最重要；
 2. 选取其中 **top-m 个最关键 token**；
 3. 通过 Centered Kernel Alignment (CKA) 比较教师与学生的**注意力结构相似度**；
 4. 使学生在内部层级上**学习教师的结构关系**（即哪些词关注哪些词）。

保留教师模型在**上下文理解和注意力分布**上的结构知识

3. 跨模型语义对齐 (Optimal Transport with Importance-Scored Mass Assignment, OTIS)
 1. 对教师和学生**的最后一层隐藏状态**进行语义对齐；
 2. 使用 **Optimal Transport (最优传输)** 计算从教师分布到学生分布的最小代价；
 3. 并根据注意力权重赋予 token “重要性分数”，让重要的 token 在传输中占更大权重。

学生能学到教师**语义空间**的分布

4. 最终训练目标结合三部分： $L_{EMO} = \alpha L_{CE} + (1 - \alpha)(L_{IRA} + L_{OTIS})$

效果

Table 1: Model Performance on Classification and SentencePair Classification Tasks. "EMO" denotes our proposed framework.

Method	Classification task				SentencePair Classification task			
	Dataset	Accuracy	Precision	Recall	Dataset	Accuracy	Precision	Recall
LLM2Vec Mistral 7B SFT (Teacher)	Patent	70.0	67.7	66.1	SciTail	96.1	96.0	95.8
Bert SFT (Student)		63.1	58.7	54.4		88.1	87.7	88.8
ULD (Boizard et al., 2025)		64.8	61.4	60.9		87.0	86.4	87.8
DSKD (Zhang et al., 2024)		64.0	60.0	58.8		88.0	87.3	88.8
MinED (Wan et al., 2024)		65.0	61.6	60.8		86.9	86.1	87.5
MultilevelOT (Cui et al., 2024)		64.6	60.4	59.0		88.2	88.0	89.1
EMO		66.5	63.3	62.4		90.9	90.1	91.2
LLM2Vec Mistral 7B SFT (Teacher)	Imdb	96.6	96.6	96.6	ConTRoL-ni	63.6	62.7	62.6
Bert SFT (Student)		91.3	91.4	91.3		42.1	38.6	37.5
ULD (Boizard et al., 2025)		92.5	92.6	92.5		45.4	45.3	45.3
DSKD (Zhang et al., 2024)		93.4	93.5	93.4		42.2	41.2	39.7
MinED (Wan et al., 2024)		92.5	92.5	92.5		47.1	47.0	47.2
MultilevelOT (Cui et al., 2024)		93.3	93.4	93.3		42.5	41.4	40.1
EMO		94.2	94.3	94.2		48.6	48.2	48.1
LLM2Vec Mistral 7B SFT (Teacher)	Banking 77	93.3	93.5	93.3	Anli_r2	67.1	67.8	67.0
Bert SFT (Student)		85.7	86.4	85.7		42.7	42.6	42.6
ULD (Boizard et al., 2025)		91.4	91.9	91.4		44.8	44.7	44.7
DSKD (Zhang et al., 2024)		91.2	91.7	91.2		43.1	43.4	43.0
MinED (Wan et al., 2024)		90.0	91.2	90.0		46.4	46.6	46.4
MultilevelOT (Cui et al., 2024)		89.4	90.4	89.4		44.1	44.1	43.9
EMO		92.3	92.7	92.3		47.6	47.8	47.5

文本对齐

Overcoming Vocabulary Mismatch: Vocabulary-agnostic Teacher Guided Language Modeling

Haebin Shin, Lei Ji, Xiao Liu, Yeyun Gong

方法 VocAgnolM

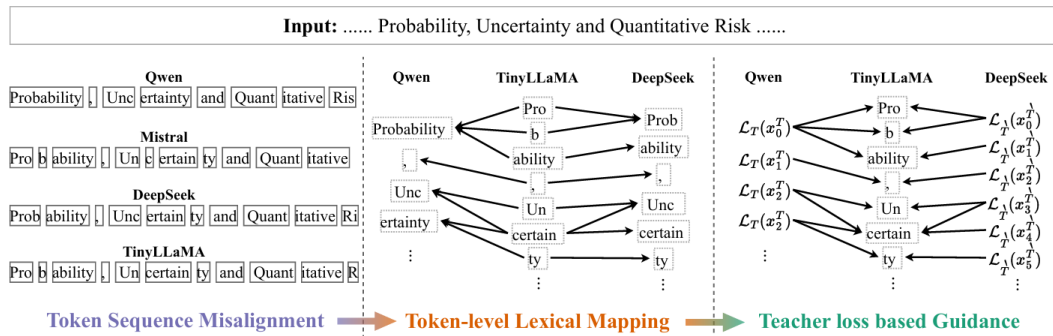


Figure 2: Overview of Vocabulary-agnostic Teacher Guided Language Modeling. *Left:* Teacher models (such as Qwen, Mistral, DeepSeek) produce token sequences that differ from those of the student model (TinyLlama), leading to misalignment. *Middle:* To address this, Token-level Lexical Mapping establishes a one-to-many mapping from each student token to corresponding teacher tokens. *Right:* To overcome logit distribution divergence, the mapped teacher token loss is utilized to guide the training of the student model.

1. Token-level Lexical Alignment (词元级词汇对齐)

- 首先使用**字符级偏移** (character-offsets) 来定位学生模型和教师模型各自的 token 在原文中的起止位置。
- 对于每一个学生 token，找出教师模型中覆盖同一原文片段 (或几乎同一片段) 的一个或多个 token。也就是“一个学生 token \leftrightarrow 多个教师 token”的映射关系 (one-to-many) 。

2. Teacher Guided Loss (教师引导损失)

- 在建立 token 映射之后，学生模型训练时会参照教师模型在对应 token（或 token 块）上的行为。
- 教师模型在其 token 输出上有一个损失（比如标准语言建模损失），学生通过映射机制，借助教师的损失/输出作为信号来训练。

效果

Overcoming Vocabulary Mismatch: Vocabulary-agnostic Teacher Guided Language Modeling													
Table 1: Performance Comparison of Student Model (<i>S</i>) Guided by Various Teacher Models. [†] Average scores for comparison with the Rho-1, following Lin et al. (2024) setup. [‡] Since SAT consists of only 32 multiple-choice questions, we report AVG score without SAT to account for abnormal cases. The best results are in bold , while second-best ones are <u>underlined</u> .													
Model	Method	GSM8K	MATH	SVAMP	ASDiv	MAWPS	TAB*	MQA	MMLU* STEM	SAT*	AVG	AVG [†] (w/o *)	AVG [‡] (w/o SAT)
TinyLlama (<i>S</i>)	-	2.7	3	10.9	17.9	20.5	12.5	13.9	16.4	21.9	13.3	11.5	12.2
TinyLlama-CPT	-	6.8	4.2	22	36.4	47.1	16.5	12.3	23.2	15.6	20.5	21.5	21.1
Teacher w/ Same Vocabulary													
Rho-1	-	7.1	5	23.5	41.2	53.8	-	18	-	-	-	24.8	-
<i>S</i> + TinyLlama-CPT	KLD	6.8	5.6	22.7	37.1	49.7	17.9	12.1	23.5	15.6	21.2	22.3	21.9
<i>S</i> + TinyLlama-CPT	Ours	7.4	4.6	21.7	37.7	48.0	16.7	13.0	22.5	25.0	21.8	22.1	21.5
<i>S</i> + Llemma	KLD	6.9	4.2	23.3	37.7	49.9	17.2	12.7	21.9	18.8	21.4	22.5	21.7
<i>S</i> + Llemma	Ours	8.1	5.2	21.9	38.1	50.1	21.0	13.9	24.0	34.4	24.1	22.9	22.8
Teacher w/ Different Vocabulary													
<i>S</i> + Mistral-ProXMath	ULD	6.0	5.4	20.9	36.4	46.7	16.7	11.2	21.1	31.2	21.7	21.1	20.6
	Ours	8.6	6.2	22.6	39.5	51.2	21.7	17.3	25.6	25.0	24.2	24.2	24.1
<i>S</i> + DeepSeekMath	ULD	6.3	4.8	22.4	36.8	46.0	16.6	12.2	22.4	31.2	22.1	21.4	20.9
	Ours	9.5	6.2	23.1	41.6	53.3	22.6	15.9	25.6	18.8	24.1	24.9	24.7
<i>S</i> + Qwen2.5-Math	ULD	5.8	3.6	21.3	36.1	47.1	18.0	11.7	22.4	34.4	22.3	20.9	20.8
	Ours	9.9	5.4	25.6	42.2	54.1	20.8	17.4	26.9	31.2	25.9	25.8	25.3
<i>S</i> + DeepSeekMath-RL	ULD	6.7	4.6	20.8	36.1	45.8	17.9	11.2	19.5	31.2	21.5	20.9	20.3
	Ours	10.8	7.2	27.3	45.9	59.6	22.6	19.1	28.1	21.9	<u>26.9</u>	<u>28.3</u>	<u>27.6</u>
<i>S</i> + Qwen2.5-Math-Inst	ULD	6.7	4.6	22.6	36.8	46.9	17.3	13.2	22.4	31.2	22.4	21.8	21.3
	Ours	11.3	7.6	28.9	46.8	60.7	22.5	20.5	30.3	40.6	29.9	29.3	28.6

Universal Cross-Tokenizer Distillation via Approximate Likelihood Matching

Benjamin Minixhofer, Ivan Vulić, Edoardo Maria Ponti

方法 AML

对齐在“相同意思的文本片段”上的概率

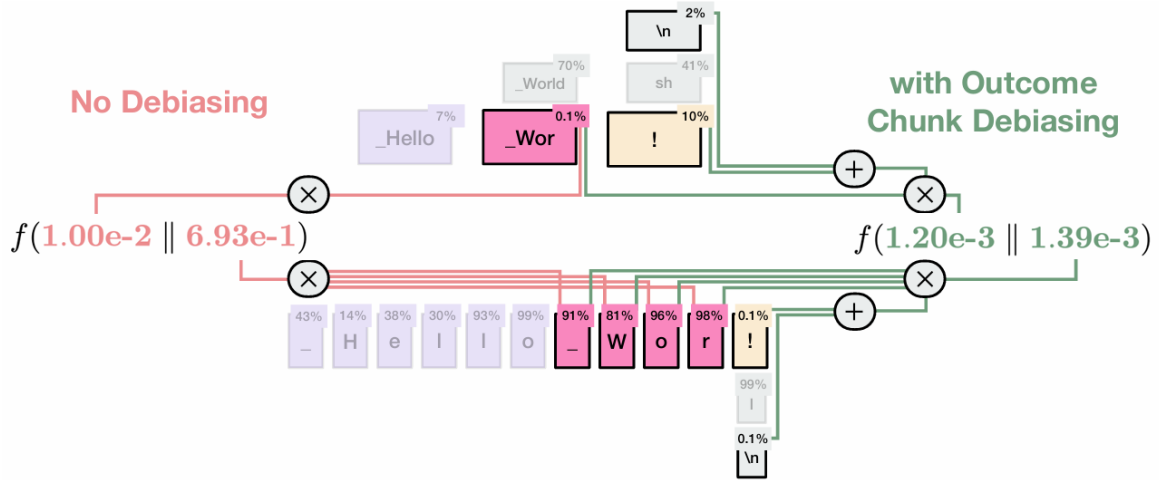


Figure 2: Outcome chunk debiasing removes tokenization bias. For example, the low probability of the subword token `_Wor` would be matched to the high-probability byte sequence `{_, W, o, r}` in naive subword \rightarrow byte transfer. We can debias by multiplying by the marginal probability of a pretoken-boundary byte occurring after the chunk. In this example, $\{\backslash n, !\} \subseteq \mathcal{B}$ and $s \notin \mathcal{B}, l \notin \mathcal{B}$ where \mathcal{B} is the set of shared pretoken-boundary bytes across the teacher and the student.

1. 识别可比 token-chunk 对齐 (chunk alignment)

- 给定输入文本 (x), 分别使用教师模型和学生模型的 tokenizer 得到两个 token 序列。
 - 在这两个 token 化序列中, 寻找“语义上对应”的子序列块 (chunks), 即教师序列中 token ($i:j$) 与学生序列中 token ($k:l$) 对应同一原文片段。公式表示为:
- $$A_c(x) = \left\{ (i, j, k, l) \in \mathbb{Z}^4 \left| \begin{array}{l} D(T_T(x)_{i:j}) = D(T_S(x)_{k:l}) = y, \\ D(T_T(x)_{i:j}) = D(T_S(x)_{k:l}) = z, \\ c(i, j, k, l) \text{ holds} \end{array} \right. \right\} \quad (\text{Alignment Indices})$$
- T_T, T_S 是教师 / 学生的 tokenization 函数, D 是解码函数。这样就建立了一组可用于比较的“chunk”对。

2. Chunk-level 概率对齐

- 对于每个 chunk 对 (i, j, k, l) , 定义教师在该老师 token 块上的概率 (teacher likelihood) $p(x, i:j) := p(T(x)_{i:j} | T(x)_{:i})$. (Chunk-Level Probability)
- 学生类似地 $p_S(x, k:l)$ 。

- 目标是 minimized 教师与学生这些 chunk 上的差异。由于 chunk 的可能数量几乎无限, 作者采用一种 **二元 (binarised) f-divergence** 的近似方式:

$$\mathcal{L}_{S,T}^{\text{ALM}}(x) = \sum_{i,j,k,l \in A_c(x)} f(p_T(x, i:j)^{\frac{1}{\tau}} || p_S(x, k:l)^{\frac{1}{\tau}}) + f(1 - p_T(x, i:j)^{\frac{1}{\tau}} || 1 - p_S(x, k:l)^{\frac{1}{\tau}}) \quad (\text{ALM Objective})$$

其中 (τ) 是温度超参数, f 是某个 f-divergence 函数。

- 这种方式不需要词表一致, 也不要求输出维度匹配, 因为它只考虑对应 chunk 的概率, 而不是整个词表维度的对齐。

3. Outcome Chunk 去偏差 (Debiasing)

- 由于不同 tokenizer 在分词偏差上不同 (“tokenization bias”), 教师-学生在 chunk 对齐中可能带有偏差。作者提出 “outcome chunk debiasing” 机制: 在计算 chunk 概率时, 额外乘以某些边界字符出现概率, 以减轻 tokenization bias 的影响。

4. 隐状态对齐可选 (Hidden-State Distillation)

- 为了进一步增强学生模型的内部语义结构学习，作者还可选地加入隐藏层状态对齐损失：将教师与学生隐藏层（hidden states）在已对齐 token/chunk 处进行距离最小化（例如用 L2 距离）：

$$\mathcal{L}_{S,T}^{\text{hidden}}(\mathbf{x}) = \sum_{l_T, l_S \in L_{T,S}} \sum_{i,j,k,l \in \mathbf{A}(\mathbf{x})} \|H_T(T_T(\mathbf{x}))_j^{l_T} - \text{proj}(H_S(T_S(\mathbf{x}))_l^{l_S})\|$$

(Hidden State Alignment Objective)

其中 $\text{proj}(\cdot)$ 是学生隐藏状态向教师维度的投影函数。

5. 总体损失与训练

- 最终训练目标可为混合蒸馏 (hybrid) 或纯蒸馏 (pure) 模式。在混合模式中，加入标准 next-token 预测任务；而在纯模式中，仅用上述 ALM + 隐状态对齐作为目标。作者还提出一种梯度权重机制 (GradMag) 来平衡不同损失组件。

效果

Table 1: Results of transferring the Gemma2 2B IT and Llama3.2 3B IT LLMs to the Qwen2 tokenizer and to byte-level tokenization. *original* denotes the original model without transfer. *ARC-C* refers to Arc-Challenge. *AGI-EN* and *AGI-ZH* refer to the English and Chinese splits of AGIEval.

Model	Tokenizer Method	Benchmark							Avg.
		PiQA	ARC-C	BoolQ	MMLU	AGI-EN	AGI-ZH	IFEval	
Gemma2 2B IT	<i>original</i>	79.6	50.4	83.8	56.9	42.1	30.7	62.5	58.0
	SFT	75.8	43.5	77.7	49.8	31.4	28.7	54.2	51.6
	DSKD	74.0	41.4	79.7	50.4	33.0	28.9	53.6	51.6
	→ Qwen2 MinED	76.7	44.3	79.6	51.8	33.0	28.8	57.1	53.0
	ALM + SFT	77.0	48.0	82.5	53.4	36.4	31.4	55.7	54.9
	ALM	76.8	49.0	82.7	53.6	38.9	31.6	53.2	55.1
	SFT	70.7	34.7	67.9	43.1	27.6	30.0	51.5	46.5
	DSKD	70.5	35.6	70.2	42.3	27.5	30.4	55.0	47.4
	→ Byte MinED	69.4	35.8	72.8	42.9	28.7	29.9	50.2	47.1
	ALM + SFT	71.5	38.2	80.5	51.0	35.6	30.4	51.9	51.3
	ALM	72.0	40.6	81.1	51.0	36.0	29.3	44.3	50.6
	<i>original</i>	76.9	43.9	78.8	62.4	36.6	40.2	76.6	59.3
	SFT	76.4	44.0	80.0	60.7	33.9	29.6	65.6	55.7
	DSKD	72.0	37.2	78.3	45.9	32.5	30.9	60.7	51.1
	→ Qwen2 MinED	77.1	44.2	82.4	60.9	35.8	29.4	71.0	57.3
Llama3.2 3B IT	ALM + SFT	77.0	44.4	79.9	61.8	37.1	32.0	74.1	58.0
	ALM	77.3	45.6	79.0	61.6	37.1	33.3	76.3	58.6
	SFT	75.2	39.8	76.8	51.5	31.5	32.6	60.8	52.6
	DSKD	71.1	36.0	65.8	48.0	32.0	30.3	57.9	48.7
	→ Byte MinED	73.2	38.7	78.6	51.1	33.1	32.3	59.6	52.4
	ALM + SFT	73.6	39.8	76.6	57.0	35.7	33.3	58.8	53.5
	ALM	73.7	40.1	76.0	55.9	35.7	33.2	49.2	52.0
	SFT	75.2	39.8	76.8	51.5	31.5	32.6	60.8	52.6
	DSKD	71.1	36.0	65.8	48.0	32.0	30.3	57.9	48.7
	→ Byte MinED	73.2	38.7	78.6	51.1	33.1	32.3	59.6	52.4

推理对齐

[CoT2Align: Cross-Chain of Thought Distillation via Optimal Transport Alignment for Language Models with Different Tokenizers](#)

Anh Duc Le, Tu Vu, Nam Le Hai, Nguyen Thi Ngoc Diep, Linh Ngo Van, Trung Le, Thien Huu Nguyen

方法 COT2ALIGN

现有方法如 Universal Logit Distillation (ULD) 与 Dual-Space Knowledge Distillation (DSKD) 已经部分解决了词典不匹配的问题，但 **往往忽视了模型的“推理能力 (reasoning capability)”** 的迁移。本文认为：教师模型不仅输出答案，而且输出推理过程

COT2ALIGN 框架:

1. Chain-of-Thought (CoT) 增强

- 在教师模型生成数据时，除了标准输出（直接答案）之外，还生成带有 CoT 的推理过程（即“思路/步骤 + 答案”）作为训练素材。
- 学生模型在蒸馏中不仅学习标准输出，还学习 CoT 输出，从而增强其推理能力。

2. Cross-CoT Alignment (跨 CoT 对齐)

- 定义两类对齐损失：
 - L_{CRC} ：将学生的标准输出与教师的 CoT 输出对齐。
 - L_{CST} ：将学生的 CoT 输出与教师的 CoT 输出对齐。
- 通过这两者，使学生模型能“模仿教师的推理过程”而不仅仅是最终答案。

3. 序列级与层级级的最优传输对齐 (Optimal Transport, OT)

- 现有 token-wise 对齐（例如 ULD 用 Wasserstein 距离对齐不同词典下的概率分布）有其局限，因为长度不同、词典不同导致 token 对齐困难。
- 本文将 OT 扩展至“序列级 (sequence-level)”和“层级级 (layer-wise)”对齐：
 - 将教师与学生在 embedding 层、最后隐藏层 (last hidden states) 作为两个序列分布进行对齐。
 - 构造成本矩阵 (cost matrix) 基于教师 / 学生 token 表示的相似度（通过投影、规范化等）计算。
 - 求解熵正则化的 OT 以获得最优运输计划 T^* ，从而定义 OT 损失 L_{OT} 。
- 这种方式不要求两者使用同一维度词典，也不要求输出长度相同，有效支持跨词典蒸馏。

4. 整体损失函数 $L = (1 - \alpha)L_{CE} + \alpha(L_{CRC} + L_{CST} + L_{OT} + L_{KD})$

- 其中 L_{CE} 是标准交叉熵监督损失， L_{KD} 是传统蒸馏损失， α 控制蒸馏 vs 监督之间比例。
- 通过上述机制，学生既从教师获取答案，也学习其推理“链条”，同时跨词典对齐其隐藏表示。

效果

Methods	Dolly	Alpaca	S-NI	Dialogue Sum	Avg
<i>Qwen1.5-1.8B → GPT2-120M</i>					
Teacher	28.23 \pm 0.13	33.76 \pm 0.33	30.32 \pm 0.26	35.37 \pm 0.26	31.92
SFT	23.78 \pm 0.38	27.20 \pm 0.23	20.45 \pm 0.23	30.25 \pm 0.22	25.42
ULD (Boizard et al., 2024)	23.77 \pm 0.41	27.50 \pm 0.50	21.37 \pm 0.34	30.23 \pm 0.10	25.72
MinED (Wan et al., 2024)	24.21 \pm 0.31	28.47 \pm 0.42	21.76 \pm 0.13	31.36 \pm 0.09	26.45
DSKD (Zhang et al., 2024b)	24.42 \pm 0.32	28.48 \pm 0.32	22.26 \pm 0.26	31.46 \pm 0.22	26.66
COT₂ALIGN	25.34\pm0.23	28.78\pm0.18	24.02\pm0.32	31.76\pm0.10	27.48\uparrow0.82
<i>Mistral-7B → TinyLLaMA-1.1B</i>					
Teacher	32.15 \pm 0.56	36.44 \pm 0.48	30.16 \pm 0.25	36.18 \pm 0.23	33.73
SFT	23.20 \pm 0.16	29.48 \pm 0.48	24.65 \pm 0.25	31.08 \pm 0.17	27.10
ULD (Boizard et al., 2024)	25.48 \pm 0.29	31.33 \pm 0.36	26.55 \pm 0.10	33.69 \pm 0.26	29.26
MinED (Wan et al., 2024)	25.54 \pm 0.59	31.82 \pm 0.33	26.13 \pm 0.23	33.31 \pm 0.16	29.20
DSKD (Zhang et al., 2024b)	26.28 \pm 0.35	32.31 \pm 0.15	26.74 \pm 0.24	33.44 \pm 0.18	29.69
COT₂ALIGN	27.41\pm0.43	33.31\pm0.49	29.77\pm0.20	35.01\pm0.20	31.38\uparrow1.69
<i>Qwen2.5-7B-Instruct → GPT2-1.5B</i>					
Teacher	28.49 \pm 0.21	35.75 \pm 0.25	32.35 \pm 0.24	35.24 \pm 0.08	32.96
SFT	21.83 \pm 0.28	27.15 \pm 0.31	23.16 \pm 0.15	30.74 \pm 0.17	25.72
ULD (Boizard et al., 2024)	24.52 \pm 0.28	29.17 \pm 0.22	24.18 \pm 0.08	32.74 \pm 0.35	27.65
MinED (Wan et al., 2024)	25.52 \pm 0.44	30.41 \pm 0.56	25.09 \pm 0.25	33.83 \pm 0.24	28.71
DSKD (Zhang et al., 2024b)	25.38 \pm 0.46	30.48 \pm 0.38	25.92 \pm 0.18	33.82 \pm 0.23	28.90
COT₂ALIGN	26.72\pm0.22	33.02\pm0.40	27.72\pm0.13	35.63\pm0.22	30.77\uparrow1.87

Table 1: Comparison of methods across different datasets. We present the $mean_{\pm std}$ values derived from experiments conducted across 5 random seeds. SFT refers to Supervised Fine-Tuning, where the student model is directly trained on the downstream dataset.