汤济玮

[个人主页] ■ 硕士■ 男 ■ 共青团员 ■ 电话/微信: 13632108098 ■ Email: tangjw24@mails.tsinghua.edu.cn

教育背景

工学硕士 ■ 清华大学 **(推免)** (985) **计算机技术** 2024.09 – 2027.06

■ 研究方向: 自然语言处理, 大语言模型, 文本压缩 ■ 硕士期间荣誉: 清华大学院级二等奖学金

工学学士 ■ 暨南大学 (211) **计算机科学与技术** 2020.09 – 2024.06 ■ 排名: **1 / 99**

■ 本科期间荣誉: 国家奖学金, 优秀毕业生, 优秀毕业论文, 优秀学生二次, 优秀学生一等奖学金二次

主要学术成果 [学术主页] (*表示同等贡献)

1. **COMI: Coarsed-to-fined Context Compression via Marginal Information Gain**
   Tang et al., *ICLR*, 2026. (**第一作者, CCF-A, 得分 8 6 6 6, 247000+ 浏览**) [文章链接][推特][量子位]
2. **GMSA: Enhancing Context Compression via Group Merging and Layer Semantic Alignment**
   Tang et al., *ACL*, 2026. (**第一作者, CCF-A, 被 SAC 推荐 Oral, 改进版本落地于淘宝首猜**) [文章链接]
3. **Read As Human: Compressing Context via Parallelizable Close Reading and Skimming**
   Tang et al., *ACL*, 2026. (**第一作者, CCF-A, 217000+ 浏览**) [文章链接][推特][量子位]
4. **Perception Compressor: A Training-Free Prompt Compression Framework in Long Context Scenarios**
   Tang et al., Findings of *NAACL*, 2025. (**第一作者, CCF-B**) [文章链接]
5. **CoS: Towards Optimal Event Scheduling via Chain-of-Scheduling**
   Zhao, **Tang** et al., *AAAI*, 2026. (**第二作者, CCF A**) [文章链接]
6. **Learning Beyond Position: Compressing Context Directly in Semantic Space**
   Tang et al. *NeurIPS, 2026.* (Under Review, **第一作者, CCF-A**)
7. **Data Distribution Matters: A Data-Centric Perspective on Context Compression for Large Language Model**
   Lv*, **Tang*** et al., *ICML*, 2026. (Under Review, **共同一作, CCF-A,** 得分 4 4 3) [文章链接]
8. ***When Hard Negatives Hurt*: Bridging the Generative-Discriminative Gap in Hard Negative Synthesis**
   Zhang*, **Tang*** et al., *KDD*, 2026. (Under Review, **共同一作, CCF-A**, avg. T: 3, avg. N: 3)

其中，COMI、RAM、Perception Compressor 和 GMSA 被 GitHub 上 300+ Stars 的项目收录。[项目链接]

实习经历

■ 2025.7 – 至今　阿里巴巴集团　未来生活实验室-大模型训练部门（基座）　研究型实习生

· **项目一：通过边际信息收益从粗到细文本压缩研究（第一作者，ICLR 2026）**

[研究问题] 当前任务相关文本压缩方法仅依据相关性保留内容，保留的信息相关但高度冗余，易误导模型产生错误输出。

[主要工作] 引入“边际信息收益”核心指标，定义为“单元与查询的相关性减去其与其他单元的冗余性”，通过该指标量化文本单元的价值；设计从粗到细的两阶段压缩流程，先基于组间边际信息收益为长文本各段落智能分配压缩预算，再依据组内边际信息收益对段落进行压缩合并，实现冗余度与关键信息保留的动态平衡。

[实验结果] 在 32×高压缩率下，COMI 使用 Qwen2-7B 模型在 NaturalQuestions 数据集上达到 49.15 的 Exact Match 分数，较次优基线提升约 25 个百分点；即使应用于原生支持 256K 长上下文的 Qwen3-4B 模型，在 16x 压缩下 F1 分数仍超越原始提示约 2 倍，同时实现超过 2 倍的端到端推理加速。

· **项目二：通过并行化的精读、略读策略压缩文本（第一作者，ACL 2026）**

[研究问题] 针对现有任务相关文本压缩方法的效率瓶颈与表示缺陷：现有方法依赖一次性装载长序列或自回归压缩导致计算效率低下；“直接删除低相关内容”易丢失关键信息、完全压缩为隐式向量丧失自然语言可解释性。

[主要工作] 从认知科学中人类阅读行为汲取灵感，将“精读核心内容、略读次要内容”的模式转化为技术方案，提出混合阅读策略的文本压缩框架 RAM；设计可并行化的精读与略读执行机制，明确精读模块保留原始文本以确保深度理解，略读模块对次相关内容进行结构化压缩以降低计算开销，同时引入对比学习机制，优化精读与略读的边界区分模型，提升场景适配性。

[实验结果] RAM 在 4×压缩率下使用 Qwen3-4B 骨干模型在 NaturalQuestions 上取得 66.59 EM 和 59.97 F1 的最优性能，显著超越所有基线；在平均长度 16K、最大 32K 的 NarrativeQA 数据集上，32×压缩时端到端延迟仅 0.20 秒，相比 LongLLMLingua(5.14 秒)和 EXIT(302.62 秒)实现大幅加速，同时保持优异的长度外推能力。

· **项目三：数据分布对大语言模型上下文压缩的影响研究（共同一作，under review at ICML 2026）**

[研究问题] 现有上下文压缩研究仅关注模型侧改进（如压缩策略设计、架构修改），忽视了数据分布本身对压缩质量的根本性影响：输入数据复杂度差异及编码器-解码器预训练知识分布错位如何系统性影响语义保留能力尚未被探索，导致压缩性能波动难以解释且跨域部署可靠性不足。
[主要工作] 首次采用数据为中心的视角，构建可控的自编码器框架：从头预训练具有精确控制数据分布的编码器与解码器（调节通用文本与逻辑密集型语料比例），通过冻结解码器并仅优化编码器参数进行压缩训练；引入信息熵量化输入复杂度，系统分析输入数据熵值与内在数据（预训练知识）分布差距对压缩质量的影响。
[实验结果] 编码器测量的输入熵与压缩质量呈显著负相关；当编码器-解码器内在分布对齐时 F1 达 81.57%，而分布差距最大时 F1 降至 69.44%。解码器内在分布起主导作用：压缩数据与解码器对齐时 F1 为 75.86%，优于与编码器对齐的 69.44%，且分布对齐的 500M-500M 模型性能优于分布错配的 500M-1B 模型。

科研经历

■链式推理：迈向最优事件规划（第二作者，AAAI 2026）

[研究问题] 针对事件规划这一 NP 难问题，解决现有方法在效率、效果、可解释性和零样本学习方面的 trade-off，实现接近理论最优的规划效果；
[主要工作] 提出新的链式推理 (CoS) 框架，通过任务分解将复杂事件规划问题拆解为可逐步解决的子问题，结合循序渐进的推理引导激活大模型规划潜力；设计 “教师 - 学生” 蒸馏体系，将搜索算法生成的高质量 CoS 推理链作为 “教师模型” 知识，通过 SFT 训练使大模型（学生模型）自主生成 CoS 推理链，兼顾规划质量与计算效率。
[实验结果] 在纽约、华盛顿和伦敦三个真实数据集上，CoS 实现了理论最优解 90% 以上的效用分数（如 Mistral-7B+CoS 在伦敦达 4.65/5.17），推理延迟控制在 5 秒内（伦敦数据集 4.79 秒），冲突率低于 20%，且在零样本跨城市测试中效用分数比基线方法高 50%。

■通过区间合并与层级式语义对齐增强文本压缩（第一作者，ACL 2026）

[研究问题] 解决软提示词压缩 autoencoder 训练中的语义占据及层级语义不对齐问题，提升文本压缩的语义保留的完整度和下游任务性能；
[主要工作] 设计区间合并模块，通过均匀区间划分与合并学习均匀的压缩表示；引入由 decoder 低层初始化的轻量级 transformer 块（层级式语义对齐模块），解决不同网络层间的语义差异问题。
[实验结果] PwC 数据集上，GMSA-AE 在 4x/8x 压缩率下上下文重构的 BLEU 分数比 ICAE-AE 高 20–30%；在 NaturalQuestions 和 2WikiMQA 等下游任务中，4x 压缩下 Qwen3-4B 达到 EM 60.38，显著优于基线，且在 32x 压缩率下仍保持最佳性能，端到端推理延迟比原始提示快约 5 倍。

■感知压缩机：一种非训练的提示词压缩框架（第一作者，NAACL 2025）

[研究问题] 解决长提示词冗余、关键信息 “lost in the middle”问题，提升长上下文场景下大模型任务表现；
[主要工作] 设计感知检索器，通过引入引导性问题解决复杂问题场景下关键信息检索难题，在 NaturalQuestions 召回实验中优化检索策略，提升 recall@1 指标；提出半指导迭代压缩方法，基于词元级别的语义重要性评估，实现关键信息 token 的精准保留与干扰性 token 的有效过滤，平衡压缩率与信息保留度；
[实验结果] 在 NaturalQuestions 数据集上，Perception Compressor 以 2.1x 压缩率（1,373 tokens）达到 78.6% 准确率，recall@1 达 72.3%（比 LongLLMLingua 高 5.2%）；在 4x 压缩下准确率达 66.0%，比次优方法高 5.9 个百分点且使用更少 token；在 LongBench 和 MuSiQue 多任务评估中均取得优越性能。

■弥合生成-判别差距：检索任务中的硬负样本合成研究（共同一作，under review at KDD 2026）

[研究问题] LLM 生成负样本有两个缺陷：1）判别无关的文本生成（生成文本过于泛化）2) 源依赖的建模捷径（模型通过识别生成痕迹而非语义进行判别）。
[主要工作] 提出 CausalNeg 框架以提升检索模型性能：1）设计 CoT 引导的反事实扰动机制，将查询需求拆解为具体信息要求，通过精准、受控地违反特定要求（如实体替换、约束违反）构建具备“有意图硬度”的可解释负样本；2）提出查询视角熵最大化正则化策略，在训练中通过最大化生成样本的相似度分布熵，最小化源身份与相似度分数间的互信息，强制模型忽略风格捷径。
[实验结果] 在 mMARCO-zh、HotpotQA、NQ 和 TQA 四个检索基准上均显著超越挖掘基线及朴素生成方法；在 TQA 数据集上 NDCG@10 提升 2.27%；定性分析证明该方法将生成负样本在嵌入空间的“纯类簇（即因生成痕迹导致的异常聚集）”比例从 24% 降低至 3%，实现了生成数据与真实语料分布的深度对齐。

个人技能

■ 技能: Python, Pytorch, Latex, Linux, SSH；语言: 普通话（母语）, 英语（CET-6）