一篇顶会论文的诞生：从零到一的实战经验分享

吴家隆(Jialong Wu)
wujialongml@gmail.com
https://callanwu.github.io/
https://github.com/callanwu

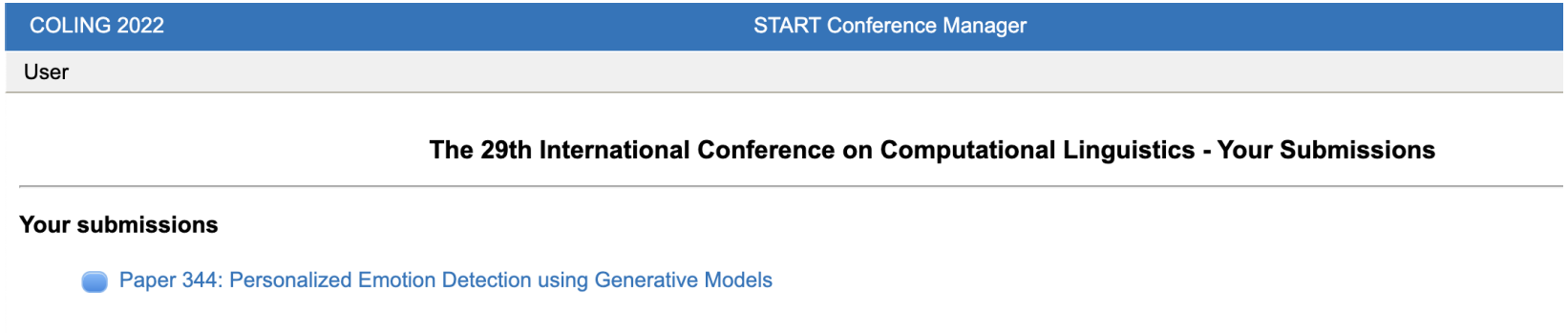
大纲

- 个人介绍
- Idea 阶段
- 写作阶段
- PR宣传阶段
- 展示阶段

个人简介

求学轨迹

在21年6月在苏州大学大二升大三的暑假（**2021年7月**）进入NLP实验室，**2022年5月**本科三年级投稿了**第一篇一作**顶会COLING2022（CCF-B）



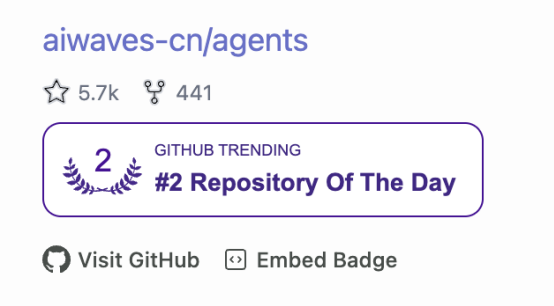
东南大学硕士阶段， **2024年2月**研究生一年级投稿两篇（共）一作ACL2025（CCF-A），**2024年5月中稿了两篇ACL**

目前（共）一作在ACL、EMNLP、COLING、AAAI都发表了论文，一次（两次）投稿后都录用了

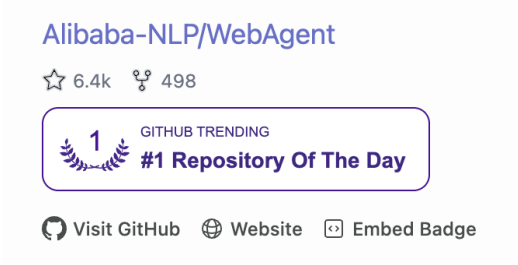
个人简介

实习经历

- 西湖大学张岳老师实验室
- AIWaves波形智能（创业公司）



- 通义实验室



个人简介

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

BERT: 2018年10月

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Colin Raffel* CRAFFEL@GMAIL.COM
Noam Shazeer* NOAM@GOOGLE.COM
Adam Roberts* ADAROB@GOOGLE.COM
Katherine Lee* KATHERINELEE@GOOGLE.COM
Sharan Narang SHARANNARANG@GOOGLE.COM
Michael Matena MMATENA@GOOGLE.COM
Yanqi Zhou YANQIZ@GOOGLE.COM
Wei Li MWEILI@GOOGLE.COM
Peter J. Liu PETERJLIU@GOOGLE.COM
Google, Mountain View, CA 94043, USA

Editor: Ivan Titov

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

Mike Lewis*, Yinhan Liu*, Naman Goyal*, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer
Facebook AI
{mikelewis,yinhanliu,naman}@fb.com

T5, BART: 2019年10月

ChatGPT: 2022年11月

LLaMA: Open and Efficient Foundation Language Models

Hugo Touvron*, Thibaut Lavril*, Gautier Izacard*, Xavier Martinet
Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal
Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin
Edouard Grave*, Guillaume Lample*

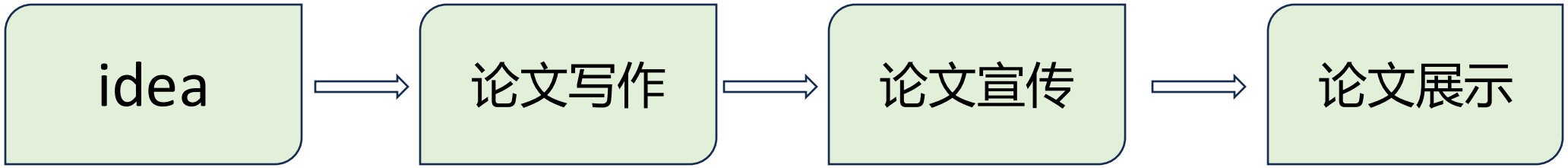
Meta AI

LLaMA: 2023年2月

研究方向

- 情感分类（2021年6月-2022年9月）
- 生成式的Parsing和图结构任务（2022年9月-2023年5月）
- Agent智能体（2023年5月至今）
- 高效LLM（2023年9月至今）

从零到一顶会论文实战



- 扎实的深度学习和NLP基础、代码能力
- 了解研究热点
- 确定研究方向
- 熟悉自己研究方向的代表工作和最新工作
- 寻找创新点

- Solid和可复现的实验
- 从Introduction到Conclusion如何写作
- 怎么画好看的图
- 怎么Rebuttal

- 宣传文案：怎么几句话概括你的工作
- 宣传渠道
- 如何扩大影响力

- 中稿焦虑
- 开会流程
- Poster
- Oral

Idea 阶段

学习资料

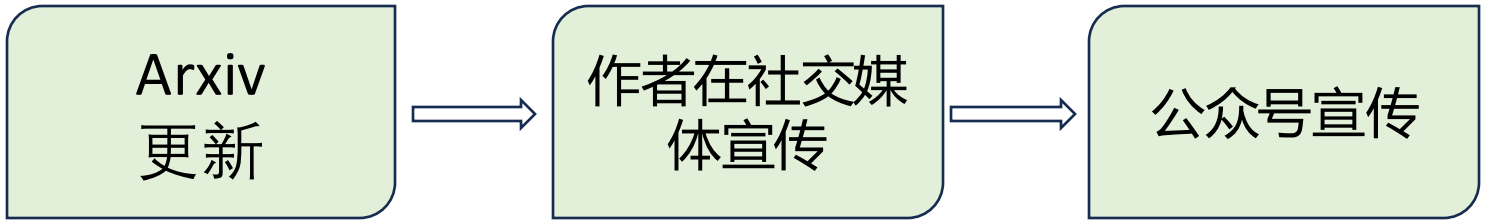
- CS224N
https://web.stanford.edu/class/cs224n/

- 跟李沐学AI
https://space.bilibili.com/1567748478/lists?sid=358497

- MIT 6.5940
https://www.youtube.com/playlist?list=PL80kAHvQbh-pT4lCkDT53zT8DKmhE0idB

Idea 阶段

科研信息获取



- **Arxiv更新**

https://arxiv.org/list/cs.CL/recent

https://papers.cool/arxiv/cs.CL

Idea 阶段

科研信息获取

- X/Twitter

最新最热门的paper追踪，建议每天空闲时间刷半小时

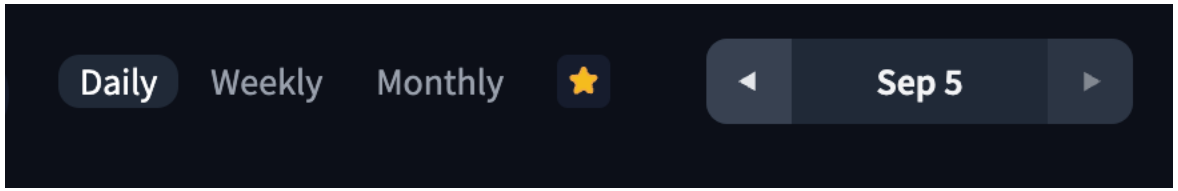
如何起号：
关注一个研究方向的学术大V，顺着评论区和关注列表一路点下去，然后相信推荐系统

Idea 阶段

科研信息获取

- 🤗HuggingFace Daily Paper

https://huggingface.co/papers/



Idea 阶段

科研信息获取

- GitHub

关注一些你所在领域的开源项目的比较活跃作者GitHub，首页会推送他们的Star和Fork

- 小红书/ 公众号/ 知乎 (网速依次递减)

机器之心，量子位，新智元

Idea 阶段

研究方向

- Follow组里研究方向
- 自己去找感兴趣的

热点 -> 竞争力大，不好中，并且容易被别人手速快抢了，并且可能资源消耗大

冷门 -> 竞争力小，做的人不多，不消耗资源，但是影响力小，找工作/读博时机会相对小

Idea 阶段

研究方向

- 看看别人在做什么
  - 工业界

研究型实习生–大规模预训练及推理的关键技术研究
更新于 2025-07-10 | 技术类 | 北京 / 杭州

研究型实习生–视觉信息驱动的多模态搜索研究
更新于 2025-07-07 | 技术类 | 北京 / 杭州 / 上海

研究型实习生–多语言大模型低资源问题的探索与研究
更新于 2025-07-07 | 技术类 | 北京 / 杭州

研究型实习生–音频分离与生成技术研究
更新于 2025-06-24 | 技术类 | 新加坡

研究型实习生–知识增强与推理能力增强研究
更新于 2025-05-26 | 技术类 | 北京 / 杭州 / 上海

Idea 阶段

研究方向

- 看看别人在做什么
  - 学术界

https://aclanthology.org/

看每年会议不同track的投稿量和接受率

Idea 阶段

具体的Idea怎么来

- 看看别人在做什么
  - 工业界

1.1.4 基于推理 RAG / Agent 的文本处理框架研究（北京）

课题简介：随着自然语言处理技术的迅猛发展，大型预训练语言模型在各种文本处理任务中展现出了卓越的性能。然而，这些模型在处理复杂推理任务时仍面临挑战。为了提升模型的推理能力，研究者们开始探索结合检索增强生成（Retrieval-Augmented Generation，RAG）和智能代理（Agent）技术的框架，以实现更高效的文本处理。本课题针对业务落地场景，重点探索以下几个方面：

1. 研究如何通过大模型的知识边界探索使其具备自我识别能力，以确定需要检索增强的知识领域。同时，开发自我反省机制，自动生成需检索的内容并对检索结果进行校验，从而确保 RAG 在推理过程中的高效性和准确性；

2. 探索如何通过多个大模型的协调合作，根据任务的复杂性和模型的特长，动态分配任务给最适合的模型。这将有效提升系统在处理复杂问题时的推理能力和响应速度。

课题简介：本课题主要研究基于大语言模型（LLM）的复杂推理与能力自我提升，旨在深入探索其在逻辑推理、任务规划和自主学习等方面的技术潜力，包括但不限于以下方向：

1. 深层次推理能力：研究大语言模型如何结合内部或外部的逻辑规则完成归纳、演绎推理以及通用推理任务的长思维链推理能力；

2. 复杂任务规划能力：研究大语言模型对复杂任务的规划、分解与执行能力，特别是如何利用有限的监督信息快速适应并拓展到新任务；

3. 自主探索与能力自我提升：研究大语言模型如何对环境进行自主探索，自动获取有效监督信号，进而完成自我学习，实现能力自我提升。

Idea 阶段

具体的Idea怎么来

- 看看别人在做什么
  - 学术界

关注会议的Workshop

Idea 阶段

具体的Idea怎么来

LLM时代下可以定义新的任务

- 老问题，新方法
- 新问题，老方法
- 新问题，新方法

经典的Idea构造想法

- 方法的A + B
- 自己熟悉的场景+热点话题

Idea 阶段

具体的Idea怎么来

- 最重要的是Motivation!
- Motivation来自于对于过去研究方向深刻的洞察和理解

看Survey
搜索最新的预印本survey看他们的未来展望!

看论文原文（最新的和最权威最有代表性的）
看他们的Limitation

看博士论文
看这个领域有代表作的刚毕业博士的毕业论文

解决什么问题(Limitation)！怎么解决(Methods)!

论文写作

论文类型

- Benchmark
- 强分析，弱方法
- 纯方法
- 纯分析
- Survey

论文写作

论文类型

- Benchmark

WebWalker: Benchmarking LLMs in Web Traversal (ACL2025)

定义任务，定义场景！
Bet 这个能力是未来大模型是最需要重要提升的部分。

	Language	Format	Depth	Width	Hop	# Pages
Mind2Web (Deng et al., 2023)	En	Multi-choice	✗	✗	✗	100
WebArena (Zhou et al., 2024a)	En	Action	✗	✗	✗	6
AssistantBench (Yoran et al., 2024)	En	QA	✗	✓	✓	525
MMInA (Zhang et al., 2024c)	En	Action	✗	✓	✓	100
GAIA (Mialon et al., 2024)	En	QA	✗	✓	✓	-
WebWalkerQA	En&Zh	QA	✓	✓	✓	1373

Table 1: Comparison between WebWalkerQA and other benchmarks. **Depth** refers to the extent of exploration required on a given website. **Width** denotes whether answering a query necessitates multiple sources. **Hop** indicates whether multiple steps are required to complete the task. **#Pages** refers to the number of webpages involved.

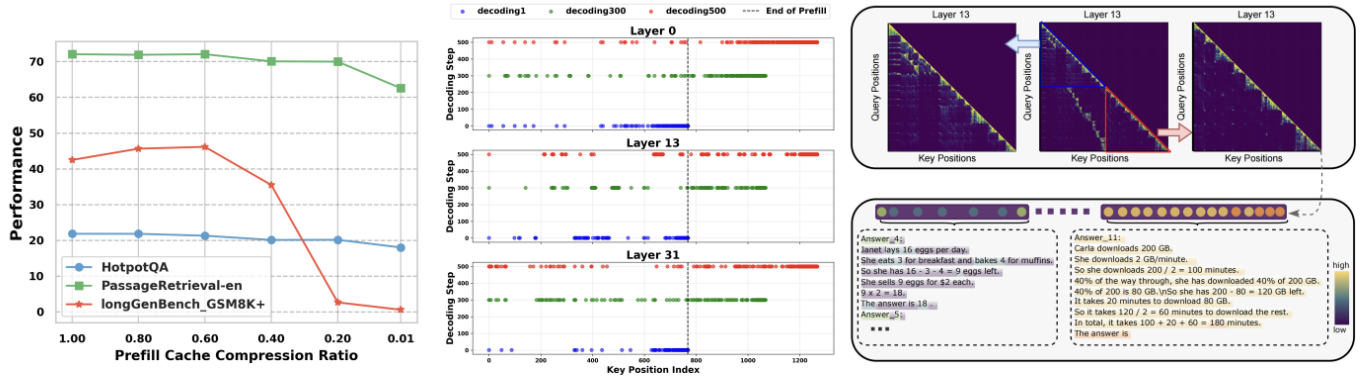
论文写作

论文类型

- 强分析，弱方法

SCOPE: Optimizing Key-Value Cache Compression in Long-context Generation (ACL2025)

Solid的前期观察实验



(a) (b) (c)

Figure 2: (a) Performances across various compression ratios during the prefill phase on three tasks under the full decoding cache condition. (b) Position distribution of the heavy hitters, selected by top 15% attention scores, at decoding steps 1, 300, and 500 across layers 0, 13, and 31. (c) Attention heatmaps for layer 13 of a GSM8k+ sample in LONGGENBENCH and details of the correspondence between attention scores and generated token positions. The complete case employed in the probing experiment is presented in Appendix 9.

论文写作

论文类型

- 强分析，弱方法

STAR: Constraint LoRA with Dynamic Active Learning for Data-Efficient Fine-Tuning of Large Language Models (ACL2024)

Solid的前期观察实验

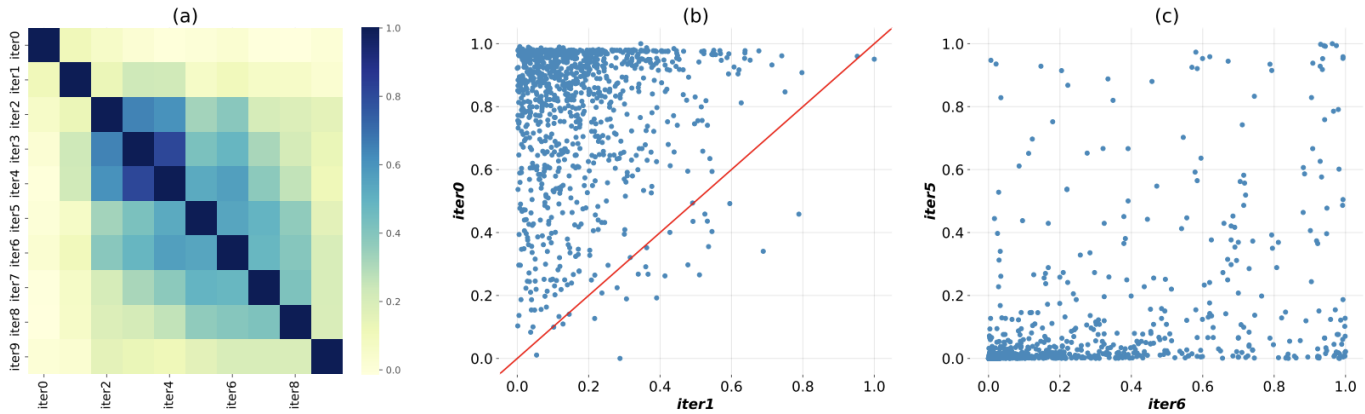


Figure 3: (a) Heatmap of correlation between prediction entropy across different iterations; (b) Scatter plot for prediction entropy between base model (Iter0) and model after first iteration (Iter1); (c) Same as (b), except values are taken from Iter5 and Iter6.

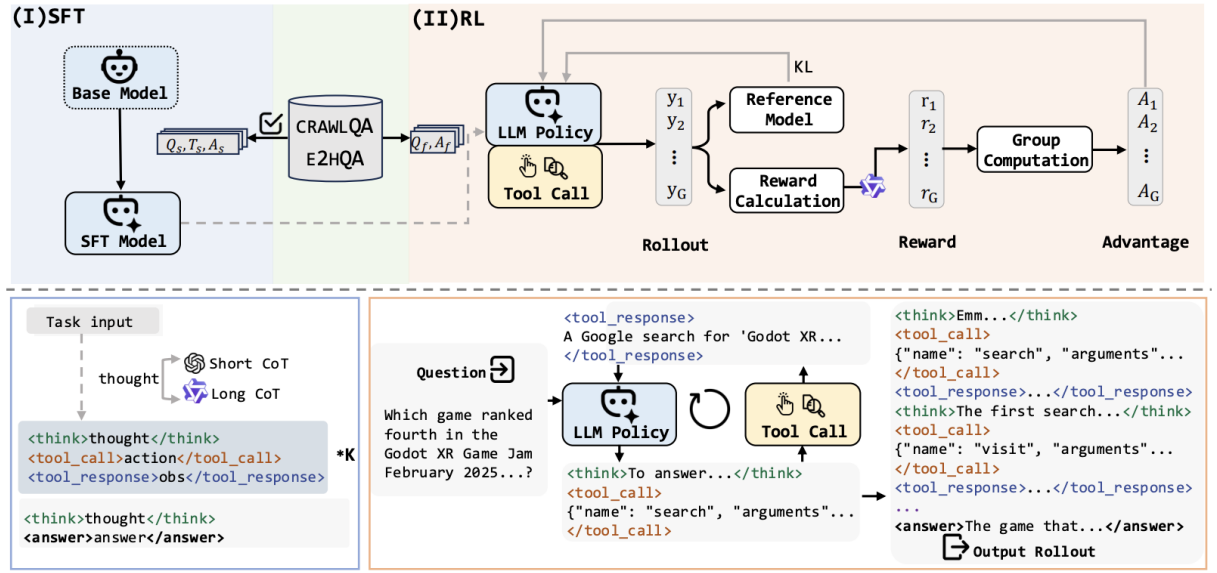
论文写作

论文类型

- 纯方法

WebDancer: Towards Autonomous Information Seeking Agency

清晰的框架



论文类型

- Benchmark
- 强分析，弱方法
- 纯方法
- 纯分析
- Survey

论文写作

方法类型

- Prompt Engineering / Context Engineering
- 微调模型

其实两者都对于工程能力有极大的要求！
大模型时代下最重要的是动手能力！

论文写作

实验部分/代码部分

- 从复现开始，一定要看高star的开源项目，避免完全从零开始
- PE的paper认真看prompt的输入输出pipeline
- 训练的paper认真看训练数据格式和修改部分的代码
- 一定要使用ai coding的工具
- Copilot/ Cursor / Claude code

https://github.com/education

论文写作

写作部分

一篇论文 = 讲好一个动听的“故事”

背景熟悉（讲得清楚相关工作） – 动机充足（明确知道Limitation，并且这个Limitation很重要） – 方法创新 （方法能够恰当得解决Limitation）– 实验有效（在对应的Benchmark上验证了方法的有效性）– 前景光明（富有insights）

机器翻译学术论文写作方法和技巧
https://nlp.csai.tsinghua.edu.cn/~ly/talks/cwmt14_tut.pdf
如何写一篇合格的NLP论文
https://zhuanlan.zhihu.com/p/58752815

必看！！！

论文写作

大模型时代下的一些特别之处

- 标题越短越fundamental，越可能是big news
- 标题里可以带emoji，prompt ChatGPT就可以

SCOPE: Optimizing Key-Value Cache Compression in Long-context Generation

WebWalker: Benchmarking LLMs in Web Traversal

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Deyu Zhou, Pengjun Xie, Fei Huang

Tongyi Lab, Alibaba Group

jialongwu@{alibaba-inc.com, seu.edu.cn}

Homepage Demo Demo Datasets Learderboard Code

- 图表一定要fancy，几个图表可能会决定一篇paper的命运，特别是在*ACL的会议

论文写作

画图

审美

- 配色：科研配色/ 糖果色
  多看别人的，然后着色器

https://colorhunt.co/

- 图标：flaticon / iconfont

https://www.iconfont.cn/

https://www.iconfont.cn/

- 画图工具

PPT / darw.io

论文写作

表格

Systems	Single-source QA			Multi-source QA			Overall
	Easy	Medium	Hard	Easy	Medium	Hard	
Close Book (No Retrieval)							
Gemini-1.5-Pro	12.50	7.86	8.33	11.25	6.43	5.00	8.08
o1-preview	16.25	10.00	9.17	7.50	10.71	6.67	9.85
Commerical Systems							
Doubao	45.00	15.00	18.33	13.75	8.57	10.00	16.76
Gemini-Search	40.00	32.14	29.17	30.00	23.57	17.50	27.94
ERNIE-4.0-8K	52.50	30.00	28.33	21.25	18.57	30.00	28.97
Kimi	77.50	41.43	40.83	26.25	26.43	22.50	37.35
Tongyi	41.25	45.00	41.67	40.00	41.43	34.17	40.73
Open-Sourced Systems							
Naive RAG	37.50	25.71	24.17	20.00	14.29	12.50	20.73
MindSearch	15.00	11.43	10.83	8.75	12.14	10.00	11.32
Avg.	37.50	24.29	23.42	19.86	18.02	16.48	-

Table 1: Performance of our proposed SCOPE using three strategies and baselines on the LONGGENBENCH benchmark with LLaMA-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3. The best results among all methods are in **bolded**. The prefill compression ratio averages around 60%.

Method	LONGGENBENCH-4K				LONGGENBENCH-8K			
	GSM8K+	MMLU+	CSQA+	Avg.	GSM8K++	MMLU++	CSQA++	Avg.
LLaMA-3.1-8B-Instruct								
Full Cache	53.26	54.40	71.67	59.78	44.44	51.01	64.92	53.46
	Decoding Compression Ratio=25.0%				Decoding Compression Ratio=12.5%			
StreamingLLM	10.78	34.15	43.50	29.48	26.98	41.26	65.33	44.53
H_2O	35.04	48.11	69.50	50.88	22.54	48.21	59.58	43.44
PyramidInfer	38.76	48.93	71.58	53.09	21.11	47.96	59.58	42.88
SCOPE *(Slide)*	**46.51**	46.62	72.50	**56.21**	**30.24**	**51.64**	**65.75**	**49.21**
SCOPE *(Adaptive)*	43.10	**50.25**	71.50	54.95	27.86	51.23	62.50	47.19
SCOPE *(Discontinuous)*	42.02	49.75	**72.67**	54.81	24.37	50.00	59.33	44.57
	Decoding Compression Ratio=12.5%				Decoding Compression Ratio=6.25%			
StreamingLLM	11.94	35.97	41.42	29.78	20.56	41.79	64.92	42.42
H_2O	26.59	45.97	65.25	45.94	21.27	45.94	55.08	40.77
PyramidInfer	28.29	46.41	61.42	45.38	19.84	45.50	55.08	40.14
SCOPE *(Slide)*	**42.56**	**50.94**	73.50	**55.67**	26.59	49.59	65.08	47.09
SCOPE *(Adaptive)*	37.29	50.19	**74.00**	53.83	**30.56**	49.94	65.92	48.80
SCOPE *(Discontinuous)*	39.85	50.06	72.92	54.27	28.41	**50.63**	**67.92**	**48.99**
Mistral-7B-Instruct-v0.3								
Full Cache	11.01	28.30	64.33	34.55	9.37	20.35	51.75	27.15
	Decoding Compression Ratio=25.0%				Decoding Compression Ratio=12.5%			
StreamingLLM	6.90	22.83	**65.25**	31.66	2.62	17.48	46.75	22.28
H_2O	7.91	26.48	60.42	31.60	5.71	16.20	40.17	20.69
PyramidInfer	10.00	24.15	62.92	32.36	5.56	16.48	40.17	20.73
SCOPE *(Slide)*	7.67	21.51	58.58	29.26	5.95	16.95	45.50	22.80
SCOPE *(Adaptive)*	11.47	**29.06**	64.50	35.01	9.76	**20.35**	**51.75**	27.29
SCOPE *(Discontinuous)*	**11.55**	**29.06**	64.50	**35.04**	**9.84**	**20.35**	**51.75**	**27.31**
	Decoding Compression Ratio=12.5%				Decoding Compression Ratio=6.25%			
StreamingLLM	6.51	19.69	57.92	28.04	3.57	17.14	46.83	22.51
H_2O	7.13	21.07	49.83	26.01	5.63	16.07	33.25	18.32
PyramidInfer	7.13	20.94	51.75	26.61	5.63	16.76	33.25	18.55
SCOPE *(Slide)*	8.84	18.68	51.75	26.42	5.71	17.08	46.17	22.99
SCOPE *(Adaptive)*	**10.93**	**29.06**	**64.50**	**34.83**	7.30	19.94	**51.75**	26.33
SCOPE *(Discontinuous)*	10.39	**29.06**	**64.50**	34.65	**8.33**	**20.19**	**51.75**	**26.76**

论文写作

Prompt

Prompts for WebWalker

The Expoloer Agent

Digging through the buttons to find quailty sources and the right information. You have access to the following tools:

{tool_descs}

Use the following format:

Question: the input question you must answer
Thought: you should always think about what to do
Action: the action to take, should be one of [{ tool_names}]
Action Input: the input to the action
Observation: the result of the action
... (this Thought/Action/Action Input/Observation can be repeated zero or more times)

Begin!

{query}

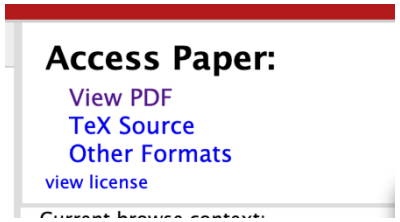
The Critic Agent

Python-style Pseudocode for SCOPE Implement

```python
# Pseudocode for Prefill and Decoding Phases with Three Strategies: Slide, Adaptive, Discontinuous

#
class CachePool:
    def __init__(self):
        self.prefill_cache = (key, value)
        self.decoding_cache = (key, value)

    def total_cache(self):
        return self.prefill_cache + self.decoding_cache

# Prefill phase
def prefill_phase(input_tokens, model, alpha1, alpha2):
    kv_cache = CachePool()
    input_query, input_key, input_value = compute_qkv(input_tokens, model)
    attention_scores = compute_attention(input_query, input_key)
    selected_key, selected_value = select_top_k_cache(attention_scores[:-alpha2], k=alpha1)
    compressed_key = [selected_key, key[-alpha2:]]
    compressed_value = [selected_value, value[-alpha2:]]
    kv_cache.prefill_cache = compressed_key, compressed_value # Update prefill_cache
    return kv_cache

# Decoding phase with SCOPE
def decoding_phase(output_tokens, model, kv_cache, beta1, beta2, strategy):
    for step in range(1, len(output_tokens)):
        token = output_tokens[step]
        current_query, current_key, current_value = compute_qkv(token, model)
        kv_cache.decoding_cache.append(current_key, current_value)
        attention_scores = compute_attention(query, kvcache.total_cache) # Attention in total_cache

        if strategy == "Slide":
            # Retain a sliding window of size decoding_window_len
            if step > max_prompt_len + beta1 + beta2:
                selected_key, selected_value = select_top_k_cache(attention_scores[alpha1+alpha2:-beta2], k=beta1)
                compressed_key = [selected_key, key[-beta2:]]
                compressed_value = [selected_value, value[-beta2:]]
                kv_cache.decoding_cache = compressed_key, compressed_value # Update decoding_cache


        elif strategy == "Adaptive":
            # Dynamically adjust beta1 based on decoding progress
            if step > max_prompt_len + beta2:
                adaptive_beta1 = beta1 * (step - beta2) // (len(output_tokens)-beta2)
                selected_key, selected_value = select_top_k_cache(attention_scores[alpha1+alpha2:-beta2], k=
    ↪ adaptive_beta1) # Use adaptive_beta1
                ... # Update decoding_cache

        elif strategy == "Discontinuous":
            # Jump to noncontinuous
            if step > max_prompt_len + beta2:
                adaptive_beta1 = beta1 * (step-beta2) // (model.max_new_token-beta2)
                jump_interval = (len(output_tokens) - beta2) // beta1  # Interval between jumps
                if step % jump_interval == 0: # Noncontinuous
                    selected_key, selected_value = select_top_k_cache(attention_scores[alpha1+alpha2:-beta2], k=
    ↪ adaptive_beta1) # Use adaptive_beta1
                    ... # Update decoding_cache

    return kv_cache
```

Figure 8: Pseudocode for SCOPE Implement.

- 学会从arxiv中下载latex源码



论文写作

Rebuttal

不涨分是常态，降低预期，做好自己；己所不欲，勿施于人

顶会rebuttal技术浅谈：站着，还把论文中了
https://zhuanlan.zhihu.com/p/602024489

浅谈学术论文rebuttal
https://zhuanlan.zhihu.com/p/104298923

论文写作

Rebuttal

大模型时代下的新问题：

1. **AI 审稿**
2. **Hack AI 审稿，图片/latex 注入review攻击（ICLR已禁止）**
3. **是否向AC举报，flag问题**

论文宣传

一定要花大量的时间做PR，才能让自己的工作有影响力！！

- **提前挂arxiv，可以让自己的paper出现在第一页**

Submissions received between (all times Eastern US)	Will be announced (all times Eastern US)	Mailed to subscribers
Monday 14:00 – Tuesday 14:00	Tuesday 20:00	Tuesday night / Wednesday morning
Tuesday 14:00 – Wednesday 14:00	Wednesday 20:00	Wednesday night / Thursday morning
Wednesday 14:00 – Thursday 14:00	Thursday 20:00	Thursday night / Friday morning
Thursday 14:00 – Friday 14:00	Sunday 20:00	Sunday night / Monday morning
Friday 14:00 – Monday 14:00	Monday 20:00	Monday night / Tuesday morning

https://info.arxiv.org/help/availability.html

Local time at arxiv.org

The current local time at arxiv.org is:

Sat, 06 Sep 2025 14:05 EDT

Deadline for submissions is 14:00 (ET), Monday through Friday (excluding weekends and holidays).

About 2 days remain until the next deadline at Monday 14:00 EDT (that is Mon, 08 Sep 2025 18:00 UTC).

New submissions received before the next deadline will be announced in the mailing scheduled to begin at Monday 20:00 EDT (that is Tue, 09 Sep 2025 00:00 UTC).

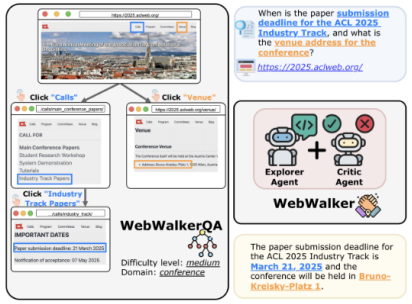
Submissions received after the next deadline will be announced in the mailing scheduled to begin at Tuesday 20:00 EDT (that is Wed, 10 Sep 2025 00:00 UTC).

See also notes on submission availability.

https://arxiv.org/localtime

论文宣传

- 利用Github ReadMe或者Github Page进行论文宣传



WebWalker: Benchmarking LLMs in Web Traversal

Jialong Wu*, Wenbiao Yin, Yong Jiang†, Zhenglin Wang, Zekun Xi, Runnan Fang,
Linhai Zhang, Yulan He, Deyu Zhou†, Pengjun Xie, Fei Huang
jialongwu@{alibaba-inc.com, seu.edu.cn}
Tongyi Lab, Alibaba Group
*Work done during internship at Tongyi Lab, Alibaba Group.

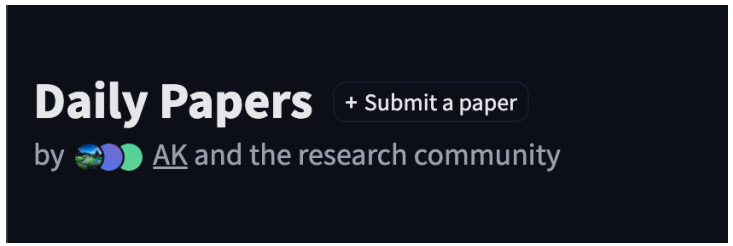
Paper | Code | Dataset | Leaderboard | Demo | Demo

https://alibaba-nlp.github.io/WebAgent/

论文宣传

- 利用HuggingFace和ModelScope等社区扩大影响力
- 上传HuggingFace Daily Paper

https://huggingface.co/papers/



- 如果是agent相关的项目一定要有demo（gradio或者streamlit）或者视频
- 如果是训练一定要有Quick Start

论文宣传

- 利用好大小公众号
- 学术社区 / 学术群聊
- 宣传物料
- 推特/小红书 推流

论文展示

- 避免中稿焦虑，相信均分，**做科研是取悦自己**

在审稿波动情况下，好的工作永远不会埋没

- 不要相信小红书的投票，存在严重的幸存者偏差
- 每个投稿周期结束后放松一段时间，去旅游或者出去玩玩
- 可以利用各种bug提前知道结果

论文展示

- **Poster**

DINER: Debiasing Aspect-based Sentiment Analysis with Multi-variable Causal Inference

Jialong Wu[♠◇] Linhai Zhang[♠◇] Deyu Zhou[♠◇✉] Guoqiang Xu[♡]

[♠]School of Computer Science and Engineering, Southeast University, Nanjing, China
[◇]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[♡]SANY Group Co., Ltd.

Introduction

Most ABSA methods solve the task as an input-output mapping problem based on high-capacity neural networks and pre-trained language models. Though remarkable progress has been made, it is demonstrated that these models are not robust in data transformation where simply reversing the polarity of the target results in over 20% drop in accuracy.

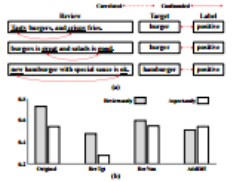


Figure 1. (a) Examples are taken from the SemEval 2014 Restaurant test set. (b) RevTgt denotes reversing the polarity of the target aspect, RevNon denotes reversing the polarity of the non-target aspect, and AddDiff denotes adding another non-target aspect with different polarity.

As shown in Figure 1 (a), over 50.0% of targets have only one kind of polarity label in the widely used SemEval 2014 Laptop and Restaurant datasets. For 83.9% and 79.6% instances in the test sets, the sentiments of the target aspect and all non-target aspects are the same. Therefore, it is easy for end-to-end neural models to learn such spurious correlations and make predictions solely based on target aspects or sentiment words describing non-target aspects.

To tackle the above challenge, we propose Debias IN AspEct and Review (DINER) based multi-variable causal inference for debiasing ABSA.

Structural Causal Model of ABSA

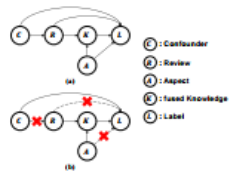


Figure 2. (a) SCM of ABSA. (b) The desired situation for ABSA, the dotted line means the causalities are blocked.

- $R \to K \leftarrow A$. The prediction of ABSA is dependent on both review R and aspect A. Therefore, a fused knowledge node K is caused by both R and A.
- $K \to L$. The label L is caused by the fused knowledge K, which is the desired causal effect of ABSA.
- $R \to L \leftarrow A$. The label L is also directly affected by review R and aspect A, where the spurious correlation comes from and should be removed.
- $C \to R$ and $C \to L$. The confounder C (the prior context knowledge) caused R and L simultaneously, where the annotation biases come from. For example, most reviews contain positive descriptions for multiple types of food, which will encourage the model to make predictions without identifying the target.

The framework of DINER

Figure 3. The framework of the proposed method.

For the $R \to L$ branch, a backdoor adjustment intervention is employed to mitigate the indirect confounding between the target sentiment words in the review and the context.

For the $A \to L$ branch, a counterfactual reasoning intervention is employed to remove the direct correlation between the target and the label.

$$TIE_{a,r} = TE_{a,r} - NDE_r + NDE_a + IE_{a,r} \quad (1)$$
$$TE_{a,r} = L_{a,r,k} - L_{a^*,r^*,k^*} \quad (2)$$
$$NDE_a = L_{a,r^*,k^*} - L_{a^*,r^*,k^*} \quad (3)$$
$$NDE_r = L_{a^*,r,k^*} - L_{a^*,r^*,k^*} \quad (4)$$
$$TIE_{a,r} = L_{a,r,k} - L_{a^*,r,k^*} - L_{a,r^*,k^*} + L_{a^*,r^*,k^*} \quad (5)$$
$$= TIE_{a,r} - NDE_a \quad (6)$$

where $TIE_{a,r}$ denotes the Total Indirect Effect (*TIE*) from A and R on L, $TE_{a,r}$ denotes the Total Effect (*TE*), NDE denotes the Natural Direct Effect (*NDE*), and $IE_{a,r}$ denotes the Interaction Effect (*IE*) between A and R.

Deconfounding the Review Branch with Backdoor Adjustment

$$L_{a,r,k} = \Psi(\zeta_a, \zeta_r, \zeta_k) \quad (7)$$

where ζ_k denotes the logit of the softmax layer, $\Psi(\cdot)$ denotes the fusion function, specially ζ_r denotes the debiased output based on R.

Consider the SCM only contains R, C, and L, C satisfies the backdoor criterion, and we can have:

$$P(L|do(R)) = \sum_c P(L|R,C)P(C) = \sum_c \frac{P(L,R|C)P(C)}{P(R|C)} \quad (8)$$

where the $do(R)$ operator denotes a causal intervention that severs the direct effect of R on L.

$$P(L|do(R=r)) \approx \bar{P}(L,R|C=c) \approx \frac{1}{K}\sum_{i=1}^{K} \bar{P}(L,R=r^k|C=c) \quad (9)$$

where $\bar{P}$ denotes the inverse weighted probability.

$$\bar{P}(L=l,R=r^k) \propto E(l,r^k;w^k) = \frac{f(l,r^k;w^k)}{g(l,r^k;w^k)} \quad (10)$$

with τ serving as a scaling factor analogous to the inverse temperature in Gibbs distributions, w^k denotes the weight parameter in the group $K=k$. The computation of logits for $P(L|do(R=r))$ is thus expressed as:

$$P(L|do(R)) = \frac{\tau}{K}\sum_{i=1}^{K} \frac{(w^k)^\top r^k}{(\|w^k\|+\epsilon)\|r^k\|} \quad (11)$$

Therefore, we model the review-specific context features C of current samples as follows:

$$C = f(r,U) = \sum_{n=1}^{N} P(u_n|r)u_n \quad (12)$$

where $P(u_n|r)$ is the classification probability of the feature r belonging to the context of class i. Now we can deblas the impact of C on R ($C \to R$) based on *TDE*. The final definition of debiased r is as follows:

$$\zeta_r = \frac{\tau}{K}\sum_{i=1}^{K} \frac{(w^k)^\top}{(\|w^k\|+\epsilon)} \left(\frac{r^k}{\|r^k\|} - \frac{r_c^k}{\|r_c^k\|} \right) \quad (13)$$

Decorrelating the Aspect Branch with Counterfactual Reasoning

The *NDE* of A on L, which represents the aspect-only bias, is calculated as follows:

$$NDE_a = L_{a,r^*,k^*} - L_{a^*,r^*,k^*} \quad (14)$$

We calculate the prediction $L_{a,r,k}$ through a model ensemble with a fusion function:

$$L_{a,r,k} = L(A=a,R=r^i,K=k) = \Psi(\zeta_a,\zeta_r,\zeta_k) = \zeta_k + \tanh(\zeta_a) + \tanh(\zeta_r) \quad (15)$$

where ζ_r is the output of the review-only branch (i.e., $R \to L$), ζ_a is the output of the aspect-only branch (i.e., $A \to L$), and ζ_k is the output of fused features branch (i.e., $K \to L$) as shown in Figure 3. *TIE* is the debiased result we used for inference.

Experimental Result

Model	Laptop Acc.	Laptop F1-score	Laptop ARS	Restaurant Acc.	Restaurant F1-score	Restaurant ARS
MemNet	-	-	16.93	-	-	21.52
GatedCNN	-	-	10.34	-	-	13.12
AttLSTM	-	-	9.87	-	-	14.64
TD-LSTM	-	-	22.57	-	-	30.18
GCN	-	-	19.91	-	-	24.73
BERT-Sent	-	-	14.70	-	-	10.89
CapsBERT	-	-	25.86	-	-	55.36
BERT-PT	-	-	53.29	-	-	59.29
GraphMerge	-	-	52.90	-	-	57.46
NADS	-	-	58.77	-	-	64.55
SENTA	67.23	-	-	77.30	-	-
PT-SENTA	74.16	-	-	80.91	-	-
ChatGPT	68.89	56.22	46.39	79.21	61.33	45.01
BERT	-	-	50.94	-	-	54.82
BERT†	70.43	66.55	49.53	78.56	69.35	57.86
DINER(BERT-based)	72.56	68.40	53.76	80.69	72.79	62.23
RoBERTa	73.57	69.26	-	79.08	72.79	-
RoBERTa†	74.96	72.16	56.27	79.26	70.47	59.96
DINER(RoBERTa-based)	76.51	73.27	59.40	82.46	76.92	64.02

Table 1. We retrained BERT†, RoBERTa† as fair baselines ensuring that comparisons are made under similar training settings, which is crucial for validating DINER's superior performance.

Case Study

Type	Example(Target Aspect: food)	Gold	Baseline	DINER
Original	The food is top notch, the service is attentive, and the atmosphere is great.	Positive	Positive ✓	Positive ✓
Rev(Tgt	The food is nasty, but the service is attentive, and the atmosphere is great.	Negative	Negative ✓	Negative ✓
Rev(Non	The food is top notch, the service is heedless, but the atmosphere is not great.	Positive	Negative ✗	Positive ✓
AddDiff	The food is top notch, the service is attentive, and the atmosphere is great, but music is too heavy, waiters is angry and staff is arrogant.	Positive	Negative ✗	Positive ✓

WebWalker: Benchmarking LLMs in Web Traversal

ACL 2025 VIENNA

Jialong Wu[1,3], Wenbiao Yin[3], Yong Jiang[3], Zhenglin Wang[1], Zekun Xi[3], Runnan Fang[3]
Linhai Zhang[2], Yulan He[2], Deyu Zhou[1], Pengjun Xie[3], Fei Huang[3]
[1]Southeast University [2]King's College London [3]Tongyi Lab, Alibaba Group
wujialongml@gmail.com

Introduction

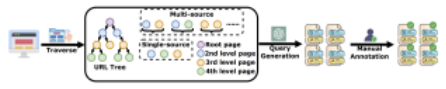
Traditional online search may **not** trace the **Deeper content** embedded within website.

How to solve it:
Interacting with the web pages and **digging through** them can effectively address **deep information seeking**.
We constrain actions to click to evaluate the agent's navigation and information-seeking capabilities.

Web Traversal Task
Given a URL root and a query, the goal of this task is to gather enough information through page traversal to ultimately answer the query.

When is the paper submission deadline for the ACL 2025 Industry Track, and what is the venue address for the conference?
https://2025.aclweb.org/

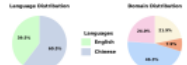
WebWalkerQA



(a) Root Official Website (b) Sublinks and Subpages (c) Synthetic QA-Pairs (d) Verified QA-Pairs

Data Generation Pipeline. We first collect root official websites. Then we mim ic human behavior by systematically clicking and collecting subpages accessible through sublinks on the root page. Using predefined rules, we leverage GPT-4o t o generate synthetic QA-pairs based on the gathered information, followed by m anual verification to ensure accuracy and relevance.

Dataset statistics on **data difficulty level**.



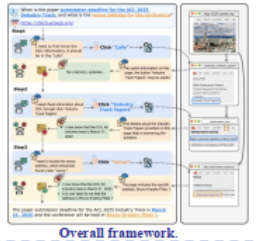
The **language and domain** distribution.

We obtain **680** question-answer pairs for WebWalkerQA.

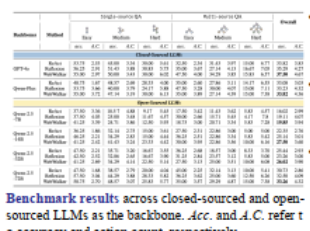
WebWalker

The explorer age nt traverses the w eb pages in **Thou ght-Action-Obse rvation** (T, A, O) paradigms.

The critic agent **update s the memory** until suff icient information is acc umulated to effectively **address the query** moti vated by pair programm ing.



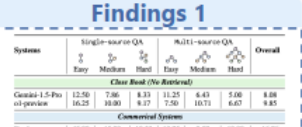
Overall framework.



Benchmark results across closed-sourced and open-sourced LLMs as the backbone. *Acc.* and *A.C.* refer t o accuracy and action count, respectively.

- Larger LLMs have e nhanced **long-range information-seekin g ability**.
- Even the best WebW alker with GPT-4o sc ores under 40%, und erscoring difficulty.
- As depth or source c ount increases, acqui ring the needed infor mation becomes har der.

Findings 1

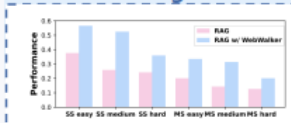


Systems	Single-source QA Easy	Single-source QA Medium	Single-source QA Hard	Multi-source QA Easy	Multi-source QA Medium	Multi-source QA Hard	Overall
Close Book (No Retrieval)							
Gemini-1.5-Pro	12.50	7.86	8.33	11.25	6.43	5.00	8.08
o1-preview	16.25	10.00	9.17	7.50	10.71	6.67	9.85
Commercial System							
Doubao	45.00	15.00	18.33	13.75	8.57	10.00	16.78
Gemini-Search	40.00	32.14	29.17	30.00	25.57	17.50	27.94
ERNIE-4.0-8K	52.50	30.00	28.33	21.25	18.57	30.00	28.97
Kimi	77.50	41.43	40.83	26.25	26.43	22.50	37.35
Tongyi	41.25	45.00	41.67	40.00	41.43	34.17	40.73
Open-Sourced System							
Naive RAG	37.50	25.71	24.17	20.00	14.29	12.50	20.73
MindSearch	15.00	11.43	10.83	8.75	12.14	10.00	11.32
Avg.	37.50	24.29	23.42	19.86	18.02	16.48	-

Results on Commercial and Open-sourced **Searched-enhanced RAG systems**.

Findings (i): *RAG systems struggle with key challenges that require effective web traversal.*

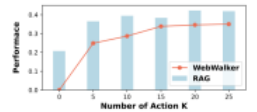
Findings 2



Performance under standard RAG and **RAG combined with WebWalker configurations**.

Findings (ii): *WebWalker can be a module in agentic RAG system, enabling vertical exploration.*

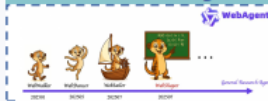
Findings 3



Overall performance on WebWalker and RAG combined WebWalker at **varying values of K**.

Findings (iii): *Scaling the process of digging through links could represent a potential direction for vertical exploration in RAG systems.*

WebAgent for Information Seeking



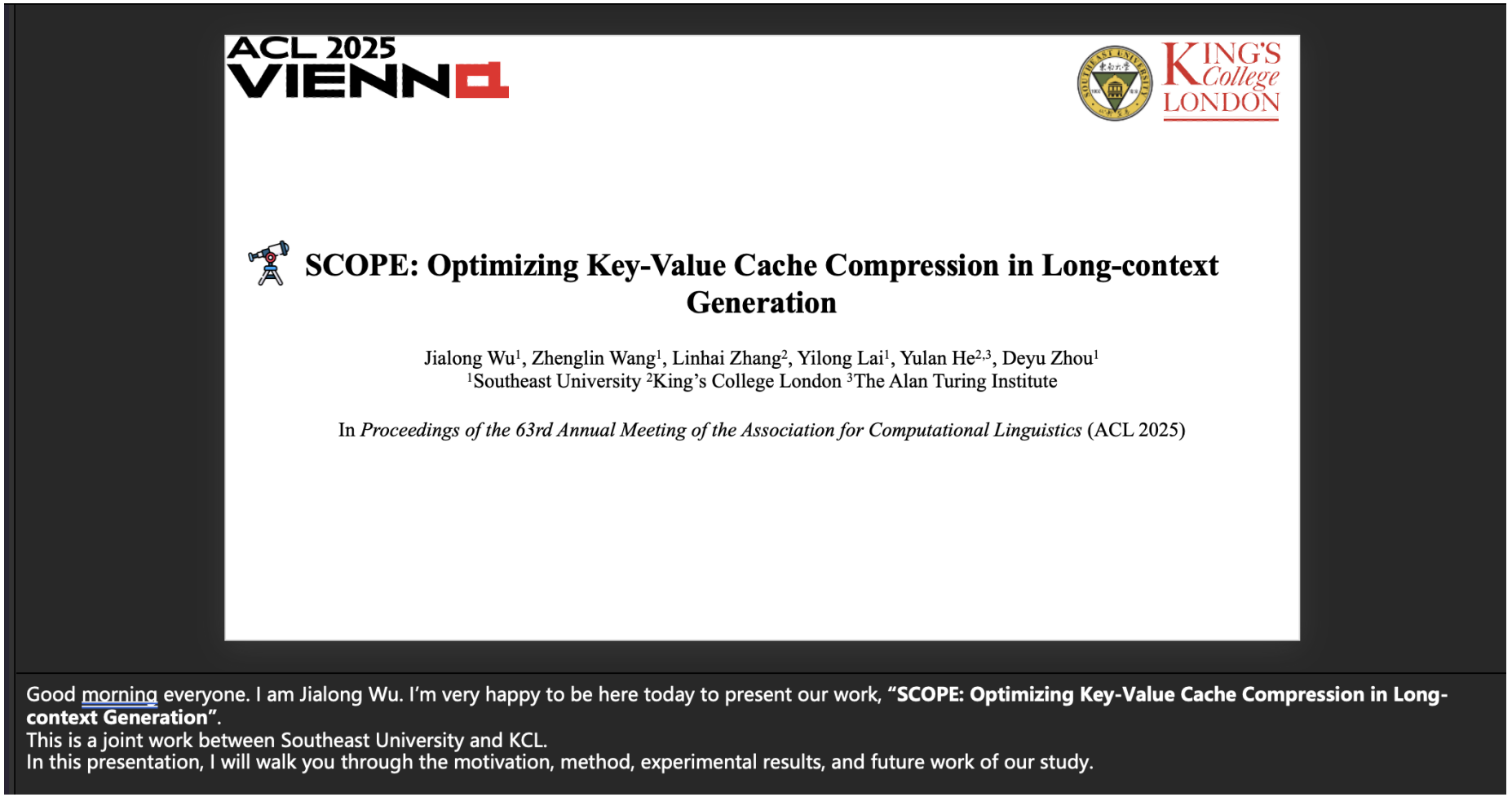
Web Agents are **autonomous** systems that perceive thei r real-world web environment, make decisions, and take actions to accomplish specific and human-like tasks.

If you like our project, feel free to give us a ⭐ on GitHub☺

Github HuggingFace arXiv Paper

论文展示

- **Oral**


ACL 2025
VIENNA
KING'S College LONDON
SCOPE: Optimizing Key-Value Cache Compression in Long-context Generation
Jialong Wu[1], Zhenglin Wang[1], Linhai Zhang[2], Yilong Lai[1], Yulan He[2,3], Deyu Zhou[1]
[1]Southeast University [2]King's College London [3]The Alan Turing Institute
In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)
Good morning everyone. I am Jialong Wu. I'm very happy to be here today to present our work, "SCOPE: Optimizing Key-Value Cache Compression in Long-context Generation".
This is a joint work between Southeast University and KCL.
In this presentation, I will walk you through the motivation, method, experimental results, and future work of our study.

结束

科研发展速度飞快，及时follow最新进展很重要，关注各类资讯
大模型时代下单打独斗可能略显乏力，要多合作
一篇有代表性的高质量的有影响力的工作胜过几篇一般的工作
质量 -> 数量

祝大家科研顺利，Paper多多，但是更重要的是生活顺利，身心愉快
抓住每一次机遇
做科研是一种取悦自己的成长

https://ccfddl.com/

结束

WebAgent for Information Seeking built by Tongyi Lab, Alibaba Group

GITHUB TRENDING
#1 Repository Of The Day

HuggingFace WebAgent | ModelScope WebAgent |

You can check the paper of WebDancer and WebWalker and WebSailor and WebShaper and WebWatcher.

💥 💥 💥 Stay tuned for more updates! We are working on building native agentic model based on the Browser and more open-domain environments!

- WebWatcher (Preprint 2025) - WebWatcher: Breaking New Frontier of Vision-Language Deep Research Agent
- WebShaper (Preprint 2025) - WebShaper: Agentically Data Synthesizing via Information-Seeking Formalization
- WebSailor (Preprint 2025) - WebSailor: Navigating Super-human Reasoning for Web Agent
- WebDancer (Preprint 2025) - WebDancer: Towards Autonomous Information Seeking Agency
- WebWalker (ACL 2025) - WebWalker: Benchmarking LLMs in Web Traversal

GITHUB TRENDING
#1 Repository Of The Day

https://github.com/Alibaba-NLP/WebAgent

If you like our project, feel free to give us a ⭐ on GitHub!

结束

吴家隆(Jialong Wu)
wujialongml@gmail.com
https://callanwu.github.io/
https://github.com/callanwu