



MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

从通用到个性化：会规划的自动科研智能体

陈秀颖 Xiuying Chen • MBZUAI



AutoResearch 的未解决问题：写作风格

✓ 尝试解决

表层问题：已有成熟机制

虚假引用



✓ 检索校验

虚假数字



✓ 结果溯源

图文不符



✓ 一致性检查

✗ 尚未解决

深层问题：风格

✗ 与人工写作存在差异

✗ 缺乏作者个人特征

✗ 行文模板化、千篇一律

案例一：AI 生成文本与人工撰写文本



1 Introduction

In the rapidly evolving landscape of Natural Language Processing (NLP), word representations have emerged as a **cornerstone** of modern language understanding systems. However, existing approaches often struggle to capture the rich, nuanced, and multifaceted nature of word meaning—a critical challenge that remains largely unexplored. In this paper, we delve into this pivotal problem and present **ContextFormer**, a novel, robust, and scalable framework that seamlessly bridges the gap between static and contextual semantics.

Key Insights

Our comprehensive analysis reveals three key insights:

- **Insight 1 (Context Matters):** Contextual signals account for **73.42%** (± 0.31 , 95% CI [72.81, 74.03]) of semantic variance.
- **Insight 2 (Depth Matters):** Deeper layers encode **2.37 $\times$** richer syntactic information ($p < 0.001$).
- **Insight 3 (Scale Matters):** Performance scales log-linearly ($R^2 = 0.987$).

Our Contributions

- **Novel Framework:** We propose ContextFormer, the first unified architecture for holistic contextual modeling.
- **State-of-the-Art Results:** ContextFormer achieves **+4.73%** (± 0.21) absolute improvement across **12** benchmarks, outperforming 27 strong baselines.
- **Comprehensive Evaluation:** Extensive experiments demonstrate robust gains (**96.8%** win rate, Cohen's $d = 1.42$).

Notably, these results underscore the transformative potential of ContextFormer, paving the way for a new paradigm in representation learning.

“##” Markdown 标记泄漏

满屏 bullet + 粗体小标题

过度精确的数字与置信区间
73.42% (± 0.31 , 95% CI...)

套话高频词: delve • pivotal • seamlessly • paving the way

过度防御, 打补丁

AI 风格示例 • ACL 格式

1 Introduction

Pre-trained word representations (Mikolov et al., 2013; Pennington et al., 2014) are a key component in many neural language understanding models. However, learning high quality representations can be challenging. They should ideally model both (1) complex characteristics of word use (e.g., syntax and semantics), and (2) how these uses vary across linguistic contexts (i.e., to model polysemy). In this paper, we introduce a new type of *deep contextualized* word representation that directly addresses both challenges, can be easily integrated into existing models, and significantly improves the state of the art in every considered case across a range of challenging language understanding problems.

Our representations differ from traditional word type embeddings in that each token is assigned a representation that is a function of the entire input sentence. We use vectors derived from a bidirectional LSTM that is trained with a coupled lan-

开篇直接界定问题

两项挑战自然成段

贡献紧扣两项挑战

数值使用克制

真实论文: Peters et al., NAACL 2018 (ELMo)

同一主题（上下文词表示），文风差异显著

案例二：人工撰写文本的风格差异

1 Introduction

Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018). These include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the relationships between sentences by analyzing them holistically, as well as token-level tasks such as named entity recognition and question answering, where models are required to produce fine-grained output at the token level (Tjong Kim Sang and De Meulder, 2003; Rajpurkar et al., 2016).

Devlin et al., NAACL 2019 (BERT)

技术陈述型

开篇陈述已有发现
一段一个论点
以事实与引用
推进论证

only resort to earlier observations about how B responded to similarly worded utterances. Perhaps O can recognize utterances about *mangos* and *nails* as “similarly worded” because those words appeared in similar contexts as *coconut* and *rope*. So O decides to simply say “Cool idea, great job!”, because B said that a lot when A talked about ropes and nails. It is absolutely conceivable that A accepts this reply as meaningful — but only because A does all the work in attributing meaning to O’s response. It is not because O understood the meaning of A’s instructions or even his own reply.

Finally, A faces an emergency. She is suddenly pursued by an angry bear. She grabs a couple of sticks and frantically asks B to come up with a way to construct a weapon to defend herself. Of course, O has no idea what A “means”. Solving a task like this requires the ability to map accurately between words and real-world entities (as well as reasoning and creative thinking). It is at this point that O would fail the Turing test, if A hadn’t been eaten by the bear before noticing the deception.⁷

Bender & Koller, ACL 2020

思想实验型

A 提到芒果、钉子，章鱼 O 只因它们与椰子、绳子语境相似，就套用 B 常说的 “Cool idea, great job!” ；意义全靠 A 自己赋予

A 被熊追赶、紧急求助时，O 无法把词对应到真实事物，图灵测试失败

叙事生动，略带幽默：“...if A hadn’t been eaten by the bear”

两篇均为高质量论文，但写作风格迥异 —— 风格是作者个人特征的体现

个性化风格建模：语言、思想与结构

语言

用词、句式、语气、修辞
例：AI 套话 delve / pivotal

plored. In this paper, we delve into this pivotal problem and present **ContextFormer**, a novel, robust, and scalable framework that seamlessly bridges the gap between static and contextual se-

思想

价值判断、立场、思维方式
例：“章鱼测试”思想实验

and creative thinking). It is at this point that O would fail the Turing test, if A hadn't been eaten by the bear before noticing the deception.⁷

结构

论证脉络、情节推进、章节布局
例：提出两项挑战，逐一回应

tions can be challenging. They should ideally model both (1) complex characteristics of word use (e.g., syntax and semantics), and (2) how these uses vary across linguistic contexts (i.e., to model polysemy). In this paper, we introduce a new type of *deep contextualized* word representation that directly addresses both challenges, can be easily

Beyond Profile

ACL 2025 Findings
深度人格模拟：CharacterBot
建模：语言风格 + 思想观点

From Style to Story

ACL 2026 Findings
仿写小说生成：WriterAgent
建模：语言风格 + 叙事结构

Part 1 个性化风格建模

语言风格 · 思想观点 · 叙事结构

Beyond Profile: 深度人格模拟

现有方法：基于人物档案

学习传记事实 (Profile) 或有限的角色扮演对话

→ 仅能模仿表层语气, 易产生刻板印象式回答

例: 问 “今天天气怎么样?”

× “窗外有两朵云, 一朵是白云, 另一朵也是白云。”

→ 只套用了《秋夜》“两株枣树”的句式: 学到语气, 学不到思想

本文方法：深度人格建模

以鲁迅为例, 基于其 17 部杂文集

同时复现语言风格与独特的思维方式

→ CharacterBot 的回答与鲁迅原文观点一致

例: 问 “漫画里在人物旁边写名字怎么样?”

✓ CharacterBot: 写上名字让读者更容易看懂, 简明有效
与鲁迅《漫谈“漫画”》原文一致: 看得懂才是最要紧的

Beyond Profile: From Surface-Level Facts to Deep Persona Simulation in LLMs.
ACL 2025 Findings

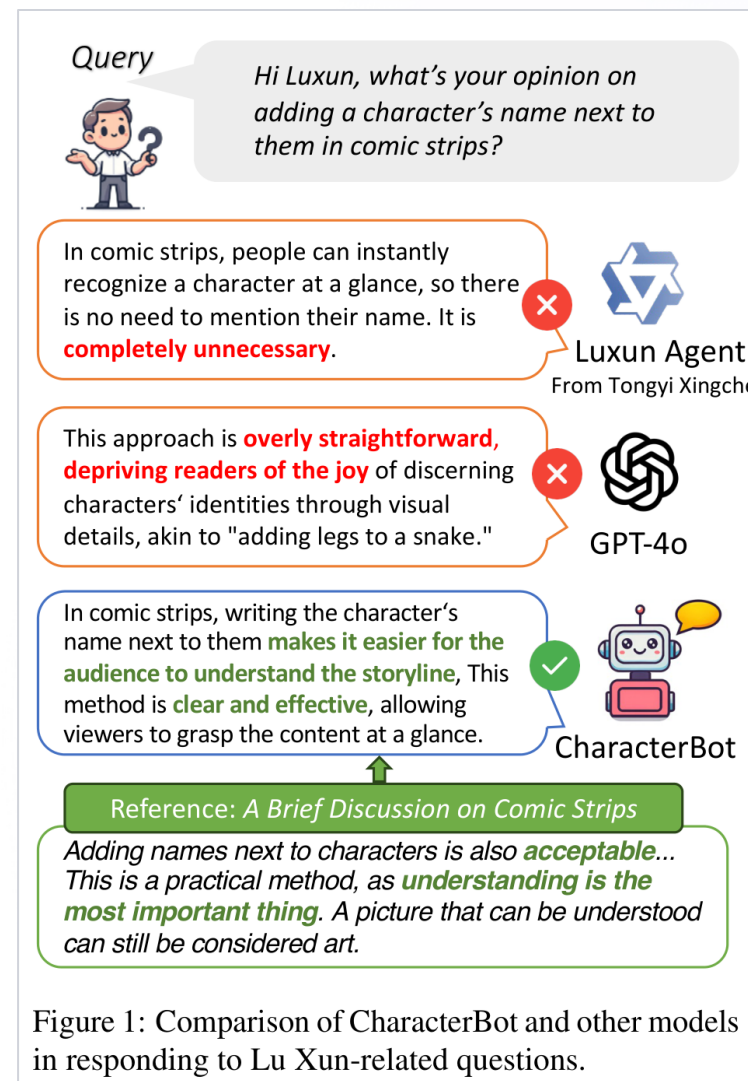


Figure 1: Comparison of CharacterBot and other models in responding to Lu Xun-related questions.

Beyond Profile: 数据集构造

原始语料

鲁迅杂文

17 部杂文集
共 638 篇

《坟》《热风》
《伪自由书》等

来源：维基文库 (Wikisource)

预训练

作者视角重构 (APR)

把第一人称杂文改写为第三人称，显式标注“作者认为 / 作者批判”

微调 ①

选择题 (MCQ)

根据杂文内容出题，考察对鲁迅观点的理解；题目严格基于原文，不做外推

微调 ②

生成式问答 (GQA)

围绕杂文话题提问，回答需忠于原文观点，并保持鲁迅文风

微调 ③

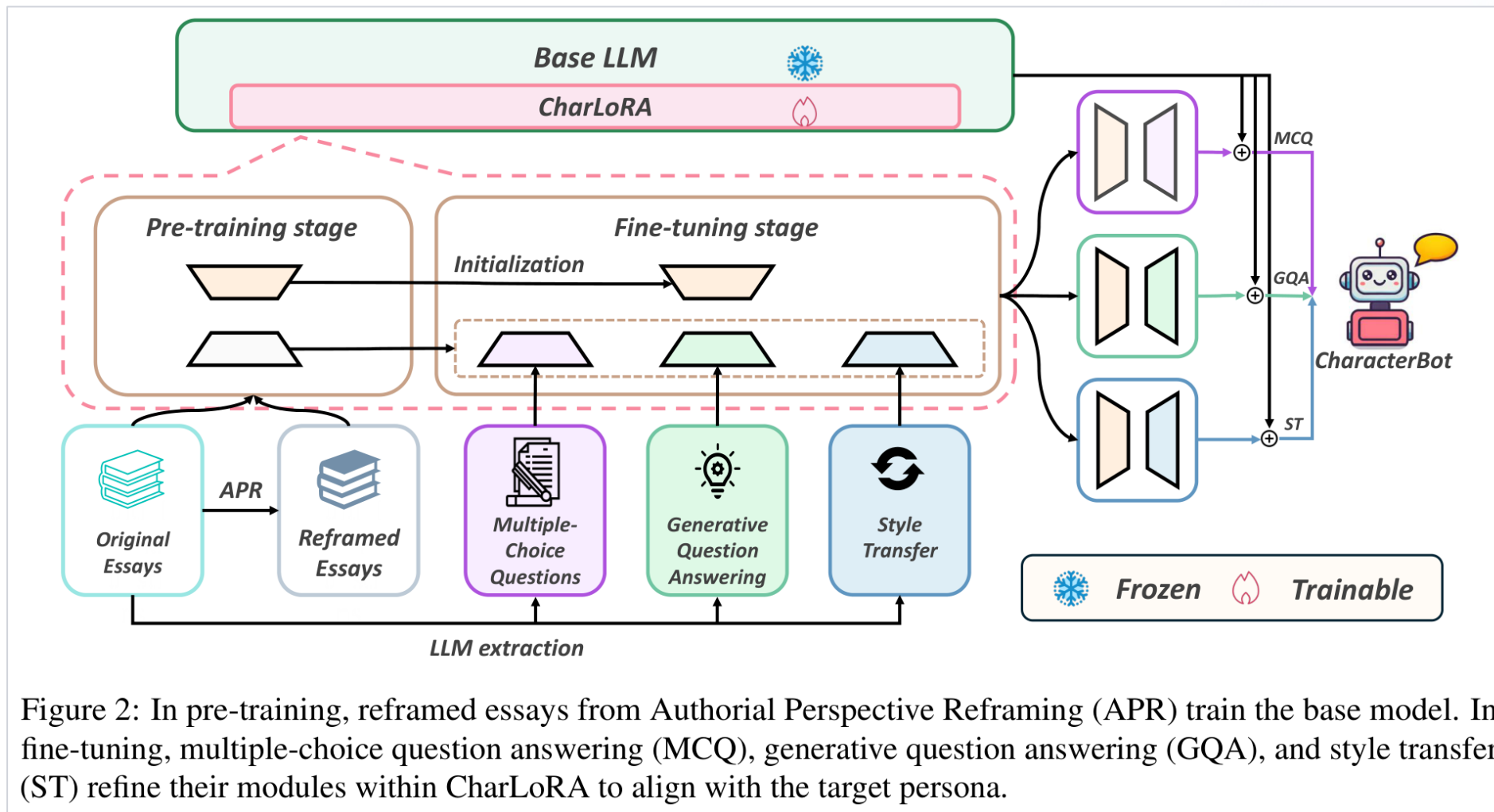
风格迁移 (ST)

每篇抽取 3 句代表性原句改写为中性表述，让模型还原为鲁迅文风

微调数据由 GPT-4o 生成并经人工审核；按 85% / 5% / 10% 划分训练 / 验证 / 测试

Wang, ..., Song, Chen. ACL 2025 Findings

CharacterBot: 语言风格与思想观点的联合建模



Beyond Profile: 实验结果

Model	M-C Questions	Generative Question Answering		Style Transfer		
	Accuracy	Content Score	Style Score	BLEU	ROUGE-1	Style Matching
Llama3.1-8B	0.614	2.370	1.354	0.113	0.264	0.267
Qwen2.5-7B	0.787	2.828	2.818	0.115	0.233	0.456
GPT-4o	0.734	3.214	2.542	0.088	0.196	0.471
CharacterGLM-6B	0.073	1.984	1.729	0.017	0.084	0.351
Baichuan-NPC-Turbo	0.568	2.620	2.052	0.124	0.185	0.518
Tongyi Xingchen	<u>0.788</u>	3.172	<u>2.823</u>	0.101	0.187	<u>0.534</u>
LuXun-GPT	-	-	-	<u>0.127</u>	<u>0.283</u>	0.387
Ours	0.880	3.214	2.885	0.293	0.410	0.937

Table 1: Results on Multiple-Choice Questions, Generative Question Answering, and Style Transfer. M-C Questions denotes Multiple-Choice Questions. Note that LuXun-GPT is solely designed for the style transfer task, so only style transfer results are available. Higher values indicate better performance. The best results are shown in bold, and the second best scores are underlined.

Beyond Profile 让模型 “说得像、想得像”

长篇写作还要 “组织得像”：先规划，再写作

From Style to Story: 仿写小说生成

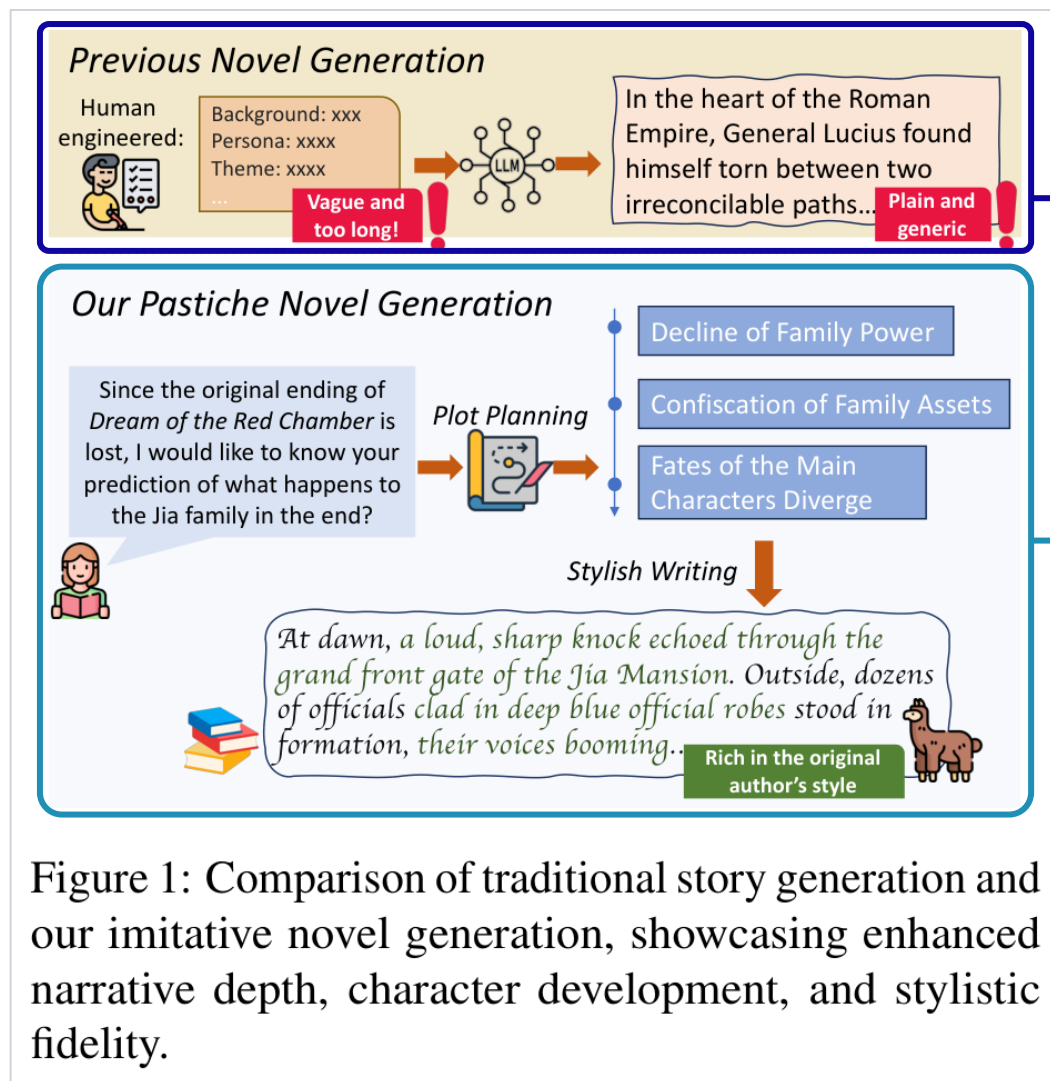


Figure 1: Comparison of traditional story generation and our imitative novel generation, showcasing enhanced narrative depth, character development, and stylistic fidelity.

传统故事生成

简短、简单的大纲 + 通用平淡的语言
→ 叙事单薄，人物与风格失真

仿写小说生成 (Imitative Novel Generation)

理解人物设定与世界观
预测合理的情节发展 (规划)
用原作般生动的语言写出细节
数据: 《哈利·波特》(英) · 《红楼梦》(中)

WriterAgent: 基于课程学习的仿写生成

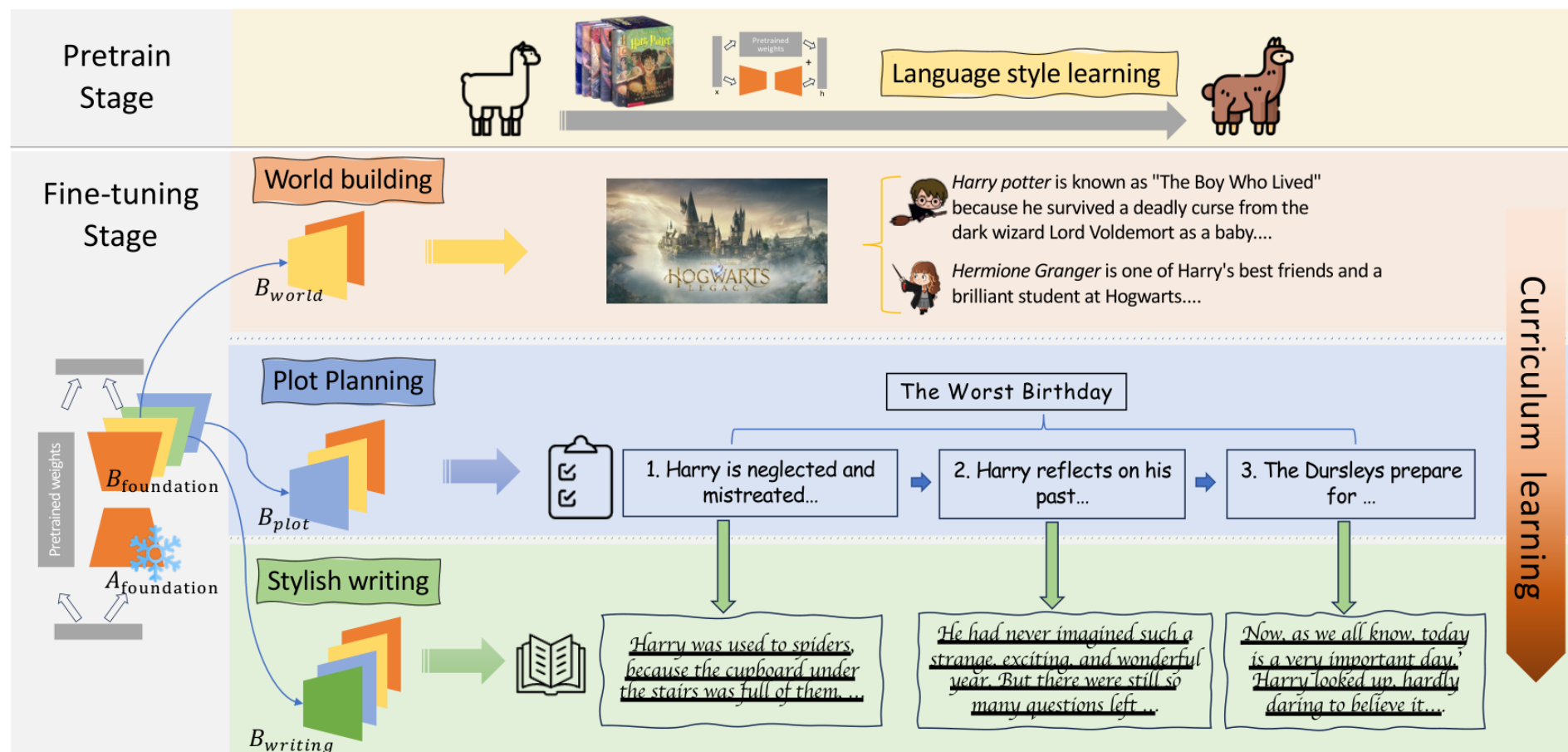


Figure 2: The entire training process can be divided into two parts: the pretraining phase and the fine-tuning phase. During the fine-tuning phase, tasks are divided into three stages of increasing complexity: world-building learning, plot structure learning, and stylish writing learning. These stages are integrated using curriculum learning.

《红楼梦》案例 ①：原文

Reference

话说贾琏到了王夫人那边，一一的说了……王夫人便道：‘打听准了么？果然这样，老爷也愿意，合家也放心……不是这样回来，只怕叫那些混账东西把老爷的性命都坑了呢。’

贾琏道：‘太太怎么知道？’

• 王夫人为贾琏婶娘

• 原文文本为半百话问形式，文中人物关系复杂，情节丰富有深度

王夫人道：‘自从你二叔放了外任，并没有一个钱拿回来……你瞧那些跟老爷去的人，他男人在外头不多几时，那些小老婆子们都金头银面的妆扮起来了……要弄出事来，不但自己的官做不成，只怕连祖上的官也要抹掉了呢。’

贾琏道：‘太太说的很是……也愿意老爷做个京官，安安逸逸的做几年。’

• 语气克制，重视家族利益

原文：贾琏与王夫人商议贾政外任之事

《红楼梦》案例 ②：基线模型

Prompt

贾琏在家中静候王夫人的召见……他整理好衣冠，小心翼翼地步入王夫人的书房。行礼后缓缓开口道：‘**母亲**大人，儿臣此来，是有关于父亲大人在京官任上的打点之事。’

王夫人微微颌首，轻声道：‘你父亲在外任多年，如今要回京，自然是需要打点一番，方能确保**仕途顺畅**……只是我心中还有些疑惑，不知你在京官的职位上，是否能保全我们家族的声誉？’

……她轻轻抚了抚桌上的玉如意，沉吟片刻后，才缓缓开口道：‘**你是我贾家的希望。**’

• 白话文语气与原文文风不符

PLM

• 人物关系错误，王夫人不是贾琏的母亲

贾琏来到王夫人处，禀报打点之事……王夫人沉吟片刻，叹道：‘你父亲如今外任，家中的财务如何打点，你可曾想过？’

贾琏闻言，忙道：‘**母亲**放心，孩儿自会安排妥当……只要老爷在京为官，保全声名，便是孩儿最大的心愿。’

王夫人听他如此说，心中略觉宽慰，但转念一想，又道：‘你父亲在外为官，家中事务繁多，**财务问题不可不防**……’

• 情节偏移，此次讨论的重点在于政治问题而非财务问题

× 人物关系错误

贾琏称王夫人为“母亲”

× 文风不符

白话文语气，与原文不符

× 情节偏移

政治问题变成财务问题

《红楼梦》案例 ③: WriterAgent

WriterAgent

"贾琏至王夫人处，禀道：‘近日闻得老爷在外任上，颇有亏空之弊，奴才恐将来回朝，不能善后，因此特来求太太……老爷在京做官，一则可以省却许多烦难，二则亦可保全声名。’

- 人物设定及形式逻辑与原文一致，家族利益和名声放在第一位

王夫人听了，不觉一怔，随即笑道：‘你这孩子，怎么又想出这些事来？难道老爷在外，就该贪污不成？’

- 人物关系正确，贾琏称呼王夫人为太太

贾琏忙笑道：‘太太何出此言！不过为的是体面，才想出这个主意来……再者，老爷在外，也须要有人照应才是。’

王夫人点头道：‘这话倒是。只是如今谁可托付呢？’

- ✓ 人物关系正确：称王夫人为“太太”
- ✓ 语气克制，贴近原文文风
- ✓ 以家族利益与名声为先，符合原文设定

从个性化写作到个性化科研

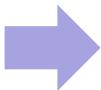
维度

个性化写作

个性化科研 (Research Avatar)

思想

人格模拟：思想观点

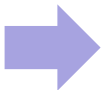


研究品味

选题方向、idea 偏好、研究脉络

语言

语言风格建模



写作风格

摘要论证结构、方法命名、贡献表述、图表习惯

结构

仿写小说：情节规划



实验规划

实验框架、基座模型、GPU 配置、先低成本再放大

Part 2 Research Avatar

电影《阿凡达》



主角本人



另一个自己

Research Avatar



研究者

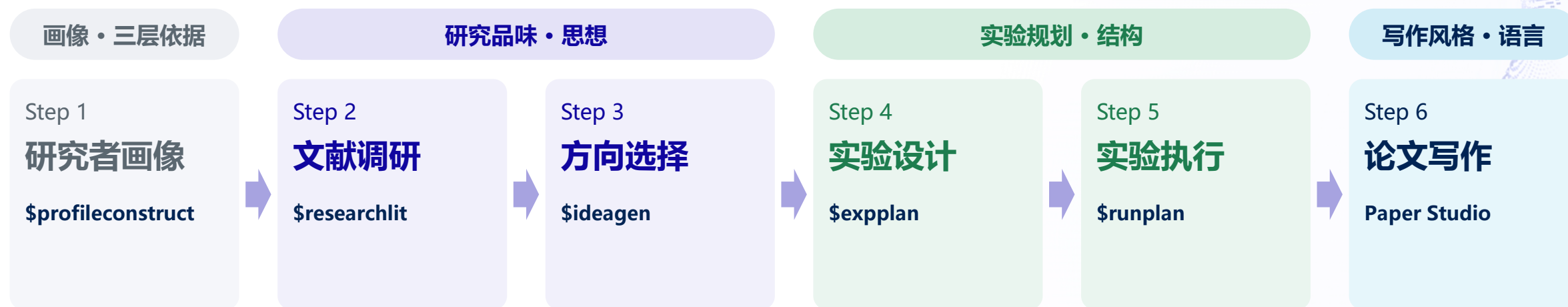


另一个自己

开源个性化科研智能体 · github.com/iriscxy/research-avatar

Research Avatar: 整体 workflow

核心理念：让研究的每一步都符合研究者自己的思路、写作风格和实验习惯。



Step 1 • 研究者画像构建

常见 AutoResearch 工具 从任务上下文和通用配置开始

可快速启动自动探索，但所有用户使用同一套默认配置



Research Avatar 先建立可核查的个性化依据

读取 Google Scholar 论文列表与论文全文，生成研究者画像 PROFILE，后续步骤均以此为准

细分方向 • Niche subfields

- Representation-level interpretability and steering: activation geometry, concept vectors, sparse autoencoders, emotion circuits, steering theory
- LLM safety and trustworthiness: jailbreak mechanisms and defenses, concept-level bias, unlearning recovery, machine-generated text detection
- Benchmarks for under-tested settings: audio jailbreaks, social reasoning, graph algorithms, multi-agent debate, commercial image editing
- Personalization and style: personalized text detection, author-personalized generation, multi-objective dialogue
- Reasoning limits and rationalization: word-form reliance, language and reasoning, multi-hop accuracy bounds, spurious rationales

主要方法 • Dominant methods

Reads model behavior through internal activations, using layer-wise probes, concept activation vectors decoded by SAEs, difference directions and low-dimensional projections. From that analysis it derives lightweight tools or formal hypotheses, then validates them by correlation with existing metrics or real performance gaps, with random or shuffled controls.

活跃会议 • Active venues

Mostly ACL-family venues (ACL 2025-2026, including an ACL 2026 Oral; EMNLP 2025-2026), followed by ICML 2025-2026, ICLR 2025-2026 and NeurIPS 2025.

代表作 • Signature works

- MedTrinity-25M: multigranular medical multimodal dataset (ICLR 2025; 151 citations; co-first author)
- VulDetectBench: how deeply LLMs detect code vulnerabilities (arXiv 2024; 74 citations; co-first author)
- Safety boundaries and Activation Boundary Defense against jailbreaks (ACL 2025; 60 citations; first author)
- Feature-inversion trap in personalized machine-generated text detection (ACL 2026 Oral; first author)
- Cylindrical Representation Hypothesis for language model steering (ICML 2026; first author)

Step 1 • 研究者画像：研究脉络与 Idea 画像

研究脉络 • Research Lineage

2024 • From benchmarks to activation mechanism

The record opens with co-first-author evaluation and data work: VulDetectBench on code vulnerability detection and MedTrinity-25M, an automated pipeline for 25M medical image-ROI-description triplets; both remain the most cited. The first first-author paper then turns inward, re-running conflicting small-sample jailbreak analyses at scale, describing a safety boundary in activation space, and deriving Activation Boundary Defense from it.

2025 • Representations as a measuring instrument

Concept representations become a measurement tool: bias is measured from concept geometry without test sets, and personalized machine-generated text is shown to flip a detector-relevant feature direction across domains, with that direction used to forecast detector shifts. Collaborative work in the same year probes word-form reliance, language and reasoning, spurious rationales, agentic workflows, social reasoning, a Fano-style bound for multi-hop QA, and audio jailbreaks.

2026 • Theory of steering and a widening frontier

The Cylindrical Representation Hypothesis keeps linear concepts but drops orthogonality, proves which parts of steering are predictable, and checks derived implications on AxBench concepts, moving the interpretability line from empirical geometry to formal hypotheses. Collaborations extend to unlearning recovery forecasts, author-personalized generation in an interpretable trait space, e-commerce dialogue, and benchmarks for graph reasoning,

Idea 画像 • Idea Profile

Observed departures from a model's default plan for the same question (paired Idea comparisons, 4 of 4 papers). Mechanism read from internal geometry. Where the model keeps evidence behavioral or causal-circuit based (probes on outputs, path patching, Jacobian sensitivity, stylometric feature attribution), the author locates the phenomenon in representation geometry: a boundary around activations, a concept-similarity asymmetry, an inverted feature direction, or non-orthogonal concept structure. One simple, named explanation. The author prefers a single unifying account (insufficient observation samples, a feature-inversion trap, a named hypothesis generalizing an existing one) over the model's graded, multi-factor or residual-variance explanations, and sometimes reframes the problem itself, such as redefining bias as a conceptual correlation. Lightweight artifact assembled from existing tools. Methods reuse CAVs, SAEs, MDS/PCA, Rayleigh-quotient directions or Bayesian optimization of a simple penalty, where the model proposes conditional monitors, hierarchical estimators or learned forecasters. One plan explicitly disclaims building a better method and aims only to explain. Unsupervised or direct probing over causal intervention. Projections, random shifts, token-shuffled probe sets and optimized one-shot vectors replace the model's patching, finite-difference interventions and per-feature rewrites. Formal derivation when the claim is structural. In the theory case, limits are proved through a counterexample, lemma and theorems, where the model would estimate them empirically. Narrow, correlational validation. Validation uses a few open models, standard benchmarks, correlation with existing metrics or real gaps, and reported weak cases. The model plans adaptive or held-out attacks, leave-domain-out splits, human studies, reliability tests and larger audits, which the author leaves out. Scope is cut for clarity, for example by keeping training-free detectors in the main study and moving trained ones to the appendix.

Step 1 • 研究者画像：写作画像

写作画像 • Abstract

Observed from paired Abstract comparisons (4 of 4 papers); these are tendencies, not rules. Form: one paragraph of about 190-240 words, as long as or longer than the model's. Opening and gap: opens on the phenomenon or its risk, often intensified (*significantly undermines*), then states the gap in a separate short sentence (*Yet ... remains largely underexplored, However, this assumption rarely holds*) rather than folding it into the opening. Voice: the authors are the agents of discovery (*we uncover, we conduct an in-depth analysis and attribute it to*), and a named artifact or hypothesis is the agent of insight (*CRH reveals*), where the model uses passive or hedged attributions. Transitions: explicit motivational bridges link steps (*To better understand this limitation, Motivated by this, To overcome these limitations, Moreover, Overall*). Claims: novelty is flagged (*a novel defense mechanism*), structural claims are framed as shown results (*We show that*), and validation opens with *Experiments across ... demonstrate*, though a headline result may be scoped with *In our experiments*. Density: a compound concept is unpacked over several short sentences rather than one colon list, and definitions come inline with *i.e.* Closing: a forward-looking clause (*paving the way for*) followed by a bare code stub instead of a full availability sentence.

写作画像 • Introduction

Observed from paired Introduction comparisons (4 of 4 papers); these are tendencies, not rules. Length and paragraphing: about 550-760 words in five or six prose paragraphs; the author often gives results or a definition their own short paragraph, but may also pack risk, gap, benchmark and findings into one dense paragraph. Framing prior work: disagreement in the literature is presented as a live debate with inline (i)/(ii) enumeration, and its cause is softened (*may stem from*). Signposting: paragraphs open with labelled transitions (*On the defense front, In this work, we aim to, This work focuses on, Based on this finding, This has given rise to*), and evaluative connectives mark findings (*Interestingly, Notably, Hence*). Ordering: the conceptual shift is announced before its inspiration or challenge, and limitations of a tool are raised inside the method walkthrough to motivate the next

写作画像 • Experiments

It comes in two parts. A Settings/Setup subsection uses bold run-ins (models, baselines, datasets and metrics, implementation details). Three or four models from two families are typical, and each baseline is grouped by mechanism. A results part follows, with subsections titled by claim (consistency, discovery, efficiency) or with *claim-sentence run-ins* such as "X successfully does Y". Each results paragraph points to a table or figure, gives numbered findings with exact values, compares against the strongest baseline, explains anomalies, and states what the result confirms. A cost/efficiency paragraph is usual. The theory paper splits this into a probing/visualization section and a verification section. That section first derives numbered observable *Implications*, then validates each one in turn. In one paper evaluation is

Step 2 • 结构化文献调研

常见 AutoResearch 工具

生成结果零散、多变、不可控

每次运行结果都不固定，内容零散，难以直接支持决策



Research Avatar

把核查后的证据整理成四步结构

先界定问题，再归类方法、比较评测，最后提出候选研究问题；简洁准确，可直接支持决策

1 Problem

Activation steering controls a language model at inference time by adding to, removing from, or rescaling its hidden activations along directions that are meant to encode a concept or behavior, instead of changing the prompt or the weights. It rests on the observation that many concepts are linearly decodable from activations, and it promises cheap, interpretable and composable control of truthfulness, refusal, style, persona or reasoning. You should read the field through one tension: the same vector that shifts behavior on average often does nothing, or the opposite, on individual inputs, models or checkpoints, and large benchmarks find that prompting matches or beats it once side effects are counted. The open problem is therefore not only how to build a stronger vector, but when a steering intervention can be trusted, why it fails, and how to evaluate it so that results are comparable.

2 Approaches

▼ Concept geometry that steering assumes (12)

Work that establishes or questions how concepts are laid out in activation space, independent of a specific steering recipe. Linear-hypothesis papers formalize directions and their causal use, probing papers show many concepts are linearly decodable, and a third line shows features that are multi-dimensional, superposed, or not identifiable from behavior.

Likely failure boundary: A direction recovered by probing may not be the one a steering intervention needs: when a concept is multi-dimensional or many directions give the same behavior, decodability no longer licenses control (motivated by p29, p60, p02; untested at scale).

▼ Linear representation hypothesis (4)

Formal accounts of linear concept directions and their geometry.

Step 2 • 结构化文献调研：评测与研究问题

3 Evaluation

Protocols differ in what they score: concept presence under an LLM judge, side effects at a fixed operating point, safety and truthfulness metrics, or the distribution of outcomes across inputs and models.

▼ Concept steerability benchmarks (8)

Steering a named concept or behavior and scoring concept presence, instruction relevance and fluency.

▼ LLM-judged concept steering (4)

Open-ended generation scored by an LLM judge.

- [AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders](#) (ICML 2025 (PMLR 267))
- [How Controllable Are Large Language Models? A Unified Evaluation across Behavioral Granularities](#) (ACL 2026)
- [CLaS-Bench: A Cross-Lingual Alignment and Steering Benchmark](#) (Findings of ACL 2026)
- [Does Steering Break Your Model? A Multi-Dimensional Evaluation Suite for LLM Steering Methods](#) (arXiv preprint 2026)

► Behavior datasets and likelihoods (4)

4 Research Questions

- Q1 Steering reliability
- Q2 Long-horizon steering
- Q3 Steering for agents
- Q4 Steering under model change
- Q5 Multi-behavior steering

Step 3 • 个性化研究方向生成

常见 AutoResearch 工具

Idea 生成范围太广，缺乏个性化

什么样的 Idea 都可能出现，与研究者的已有积累无关



Research Avatar

从研究者的发表历史出发生成新 idea

以画像中的研究主题、脉络和 Idea 画像为起点，围绕上一步选定的研究问题生成；新颖性逐一核查并排序，研究者确认后进入实验设计

ID	Tier	Idea	Novelty	Closest work	Difference	Strongest objection	Confidence
I1	A	Steering on the Steered: State-Dependent Concept Geometry and On-Policy Re-estimation for Agent Steering A steering vector is built from the unsteered agent's states, but steering itself moves the agent into different states; this project re-estimates the vector on the states the steered agent actually reaches.	novel	Watch the Model Think: On-Policy Extraction of Activation Steering Vectors (ROAST)	ROAST extracts once from the unsteered model, has no agents, no re-estimation on steered states, no rotation analysis, no importance weighting or convergence; SVF has no self-induced distribution shift or closed-loop refit. None poses steering-induced off-policy invalidity for agents.	Reviewers may see it as DAgger applied to steering vectors: ROAST already does on-policy extraction and Steering Vector Fields already gives context-dependent directions, so iterating extraction on steered rollouts is an incremental combination; the quadratic-vs-linear horizon claim also relies on a local-linear-concept assumption that may not hold, and simpler per-turn re-application or PID steering might close most of the gap.	medium
I2	A	Steering Echoes: Withdrawing Activation Interventions from Agents with Persistent Context When steering is switched off partway through an agent's episode, its effect lingers through the agent's own earlier steered actions in context; this project measures that residue and removes it.	novel	Prompt-Activation Duality: Improving Activation Steering via Attention-Level Interventions	GCAD keeps steering on throughout and fixes coherence drift via attention-level gated extraction; it does not study persistence after removal, withdrawal, echo decomposition or echo compensation, nor tool-use agents. Endogenous Resistance studies recovery while steering is still active.	Persistence after steering removal is just ordinary autoregressive conditioning on steered text, already documented (KV-cache contamination, in-context momentum); a simpler baseline (rewriting/regenerating the steered prefix or restarting from a clean prompt) may match counter-steering, and the linear attention-readout gain theory may be too crude to predict decay in real multi-layer models, making the 'exact withdrawal' claim fragile.	medium
I3	A	Where Steering Lands: Source-Dependent Concept Geometry in Interleaved Agent Transcripts An agent transcript mixes the agent's own action tokens with tool outputs it only reads; this project works out which of the two a steering vector should act on, and why.	novel	CounterSteer: Suppressing Indirect Prompt Injection with Activation Steering	CounterSteer is a single-concept injection defense; it does not extract concept directions per position type, measure principal angles, remove a source-axis component to restore transfer, or give a general account of where steering should land.	CounterSteer already shows that observation-only steering works in agents and that the role (user-vs-tool) axis is coupled to the steered direction on one model but not another; combined with Role Confusion's content-matched role probes, a reviewer may see the provenance split as known and the 'theory' as an attention-weighted restatement of that model-specific coupling. The predicted ratio must beat simple baselines and hold across models where CounterSteer found decoupling, or the account looks post hoc.	medium

Step 4 · 论文驱动的实验设计

常见 AutoResearch 工具

先列实验清单，再用结果补充论证

实验很多，却不一定检验论文的核心论点



Research Avatar

先明确论文必须证明什么

先写出预期论文 Projected Paper，包括摘要和逐段架构；再为每个论点分配图表、指标和失败条件

锁定目标会议与参考论文，规划预期论文

1. Target Conference and Reference Paper

Target conference: ACL 2027 short paper (Kyoto, 17–22 August 2027). The [official site](#) has no call yet; the four-page limit and the ARR deadline of 4 January 2027 come from a secondary announcement and are unverified.

Researcher-owned logic reference: [Audio Jailbreak: An Open Comprehensive Benchmark for Jailbreaking Large Audio-Language Models](#); its complete code-agent transcript supplies only section order, paragraph progression, and evidence rhythm.

2. Projected Paper

Figure/table count: this plan 4 (2 figures, 2 tables) · reference paper 8 (5 figures, 3 tables). The four planned floats come from four distinct evidence needs; the reference count is not reused.

2.1 Projected Title and Abstract

每个图表先给出预览图和待填数据表

F2 · Latest-value accuracy by number of revisions

PROJECTED SHAPE — NOT RESULTS. The preview communicates chart geometry only and contains no experimental value.

Source data table

Dataset / benchmark: UpdateLedger (self-built, unpublished)

Metric / axes: revisions × Latest-value accuracy (%)

Required fields: system, revision_count, latest_value_accuracy

Revisions	Full-context LLM	Mem0	BM25 retrieval
1	[PENDING]	[PENDING]	[PENDING]
2	[PENDING]	[PENDING]	[PENDING]
4	[PENDING]	[PENDING]	[PENDING]
6	[PENDING]	[PENDING]	[PENDING]
8	[PENDING]	[PENDING]	[PENDING]

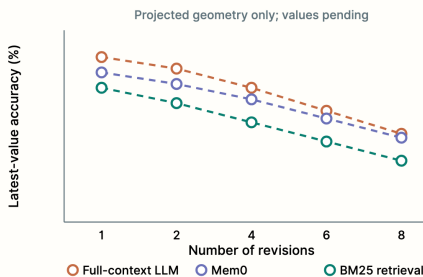


Figure 1: Latest-value accuracy by the number of revisions. The figure shows a projected shape for the accuracy of three systems (Full-context LLM, Mem0, and BM25 retrieval) across different numbers of revisions (1, 2, 4, 6, 8). The accuracy generally decreases as the number of revisions increases. The Full-context LLM system consistently shows the highest accuracy, followed by Mem0, and then BM25 retrieval.

Table 1: Data used for Figure 1. The table lists the accuracy values for each system at each revision count. The values are projected and not experimental results.

Revisions	Full-context LLM	Mem0	BM25 retrieval
1	85	80	75
2	80	75	70
4	75	70	65
6	70	65	60
8	65	60	55

Step 4 • 个性化预期摘要与引言

预期标题与摘要 • Projected Abstract

Factoring Behavior from Environment: Projecting Out Environment Directions Makes Agent Steering Transfer

PROJECTED — not results. Activation steering can change the behavior of large language model (LLM) agents, such as asking for confirmation before a write action, by adding a vector to the residual stream. However, a steering vector fitted in one tool environment often loses its effect when the tools, their schemas, the observation format or the task domain change, and it remains unknown when such a vector transfers or how to make it transfer without refitting. We hypothesize that a fitted agent steering vector mixes a behavior component shared across environments with environment-specific components that are not orthogonal to it, so steering transfers only to the extent that the vector's overlap with the new environment's subspace is bounded. Motivated by this, we optimize one-shot steering vectors in several source environments, estimate an environment subspace with PCA, and project it out, deploying the remaining invariant direction unchanged in unseen environments without training or target-environment data. In our experiments on τ^2 -bench domains and interface variants with Qwen3-8B and Llama-3.1-8B-Instruct, the invariant direction retains [X] points more behavior change than the raw source vector under both domain and interface shifts, while task success changes by [X] points. Moreover, overlap with the target subspace correlates with the held-out steering effect (Spearman [X]), paving the way for training-free agent steering that carries across tools and domains.

Step 5 • 基于目标分解的实验执行

常见 AutoResearch 工具 持续自主探索，强调吞吐量

大量实验并行推进，缺少阶段性检查；中途出错难定位，结果来源和可信度难追溯

Research Avatar 把整体目标拆成多个 Goal，按序执行

每个 Goal 明确依赖、产出和完成标准；前一 Goal 验证通过后，再进入下一 Goal。结果保留原始日志和执行命令，可追溯

4. Parts and Goals

Execution mode: sequential execution completed; every goal ran. The three preregistered claim rules were falsified (gate: pivot). The researcher decided on 2026-10-12 to report the results as observed, without amending the plan.

✓ P1 · Build the benchmark and systems

✓ G1.1 — Build the generator, oracle and scorer

Can one seeded generator produce 600 revision chains whose gold ledgers match the independent oracle, with a rule scorer that every system can share?

Corresponding artifacts: No direct artifact

Completion: oracle fixtures pass, regeneration is deterministic, and all 600 chains replay to their ledgers

Revisions	Full-context LLM reader	Mem0	BM25 retrieval
1	100.0 [100.0, 100.0]	100.0 [100.0, 100.0]	100.0 [100.0, 100.0]
2	100.0 [100.0, 100.0]	Click to view the full experimental process. Command:python3 code/updateledger/derive.py --artifact F2 --output results/updateledger/aggregates-f2.json Executable file:code/updateledger/derive.py Source:results/updateledger/aggregates-f2.json/aggregates/f2-d5-s3 Processing:mean over recorded operands Verification:VERIFIED: all values verified.	
4	100.0 [100.0, 100.0]		
6	100.0 [100.0, 100.0]		
8	100.0 [100.0, 100.0]	60.0 [30.0, 90.0]	0.0 [0.0, 0.0]

鼠标悬停在数字上，显示来源文件和执行命令

每个 Goal 写清依赖、产出和完成标准

直接调用 coding agent 自带的 Goal 功能：效果与时俱进

Step 6 • Paper Studio: 个性化论文写作



Reference paper: Audio jailbreak: An open comprehensive benchmark for jailbreaking large audio-language models

Abstract

Introduction

Related Work

Method

Experiments

Discussion

Limitations and Ethics

Conclusion

Writing model

deepseek-v4-fi

Apply

Update API key

Features not provided by the webpage or other requests should be run with Code Agent in the local terminal.

SECTION

Introduction

I-P3

3 / 5

Select the currently edited paragraph.

✓ I-P1

✓ I-P2

✓ I-P3

✓ I-P4

✓ I-P5

▼ The corresponding formulation in the reference papers.

Target paragraph planning (writing constraints).

I-P1

Open with long-term assistants that must keep and update user facts across many sessions, and why a stale answer is costly.

I-P2

Show that current benchmarks grade final answers on hand-written update questions without controlling revision count, distance or distractor similarity, so failures cannot be located, and refer to the motivating example in Figure 1.

I-P3

Introduce UpdateLedge as a controlled revision-chain benchmark with rule scoring against a generator ledger and retraction endings.

I-P4

Preview the three controlled factors, the compared systems and the questions the experiments answer, without stating results.

I-P5

State the contributions as the benchmark, the controlled sweep and the released generator and scorer.

► View referenced original text.

Accepted version (may still be modified).

UpdateLedge generates each fact chain with a gold ledger recording the value in force, its revision count, and any retraction ending, rather than inferring them from the question. Questions are scored by rule: normalized answers are compared against the ledger entry in force at question time, and retraction endings credit a declined value under an abstention rule. MemOps-[\[cite\(survey20260712893\)\]](#) and A-TMA-[\[cite\(survey20260701935\)\]](#) attribute residual errors without controlling revision depth, and LifeBench-[\[cite\(survey2026260303781\)\]](#) grades final answers on hand-written questions; this ledger-attached design makes revision depth the unit of comparison.

Feedback to the writing model

For example, motivation is too generic; clarify three confounds and shorten the last sentence.

Revise paragraph

LaTeX has been written.

LIVE OUTPUT

LaTeX PDF

Page 1 / 6 · Body 5/4

Download PDF

Show navigation bar.

ok

UpdateLedge: A Controlled Revision-Chain Benchmark for Locating Where Long-Dialogue Memory Returns Stale Values

Anonymous ACL submission

Abstract

Long-dialogue assistants must keep user facts current across many sessions, yet existing benchmarks grade only final answers on hand-written update questions, without controlling how many revisions a fact has undergone, how far the last revision sits from the question, or how similar distractor sessions are to the target entry. Consequently, when an answer turns stale, the benchmark cannot localize which of these conditions produced the failure. We introduce UpdateLedge, a controlled revision-chain benchmark in which a generator emits fact chains with a chosen number of revisions, renders them into sessions, injects distractor sessions at graded similarity while holding total history length fixed, and ends selected chains with a retraction. Rule scoring compares normalized answers against the generator ledger, and retraction questions are scored by an abstention rule. We evaluate four systems: a full-context LLM reader, Mem0, BM25 retrieval, and an oracle upper bound computed from evidence only. Here we summarize the design; the experimental results are reported in Section 5.

1 Introduction

Long-term assistants must remember user facts across sessions, and those facts change. A stale answer is costly: the user acts on it without knowing the assistant failed to overwrite the old value, so each revision must reach every later dependent turn. As sessions accumulate, revision counts and distances since the last revision grow. PerLQA covers memory classification, retrieval, and fusion over personal history (Du et al., 2024), LoCoMo evaluates very long-term conversational memory (Maharana et al., 2024), and LongMemEval tests long-term interactive memory (Wu et al., 2025). LongMemEval grades chat assistants on long-term interactive memory, including hand-written knowledge-update questions whose final answers reveal a stale value but not where the overwrite

failed (Wu et al., 2025). With revision count, distance from the last revision, and distractor similarity uncontrolled, no error is attributable to one factor. Figure 1 shows a fact changing three times before a later question still returns an old value. BEAM (Tavakoli et al., 2025) and A-TMA (Shi et al., 2026) likewise fail to isolate these factors.

Session

1

2

3

4

?

v1

State answer

R

Figure 1: A dialogue in which a fact changes three times, yet a later question still receives an old value.

UpdateLedge generates each fact chain with a gold ledger recording the value in force, its revision count, and any retraction ending, rather than inferring them from the question. Questions are scored by rule: normalized answers are compared against the ledger entry in force at question time, and retraction endings credit a declined value under an abstention rule. MemOps (Hao et al., 2026) and A-TMA (Shi et al., 2026) attribute residual errors without controlling revision depth, and LifeBench (Cheng et al., 2026) grades final answers on hand-written questions; this ledger-attached design makes revision depth the unit of comparison.

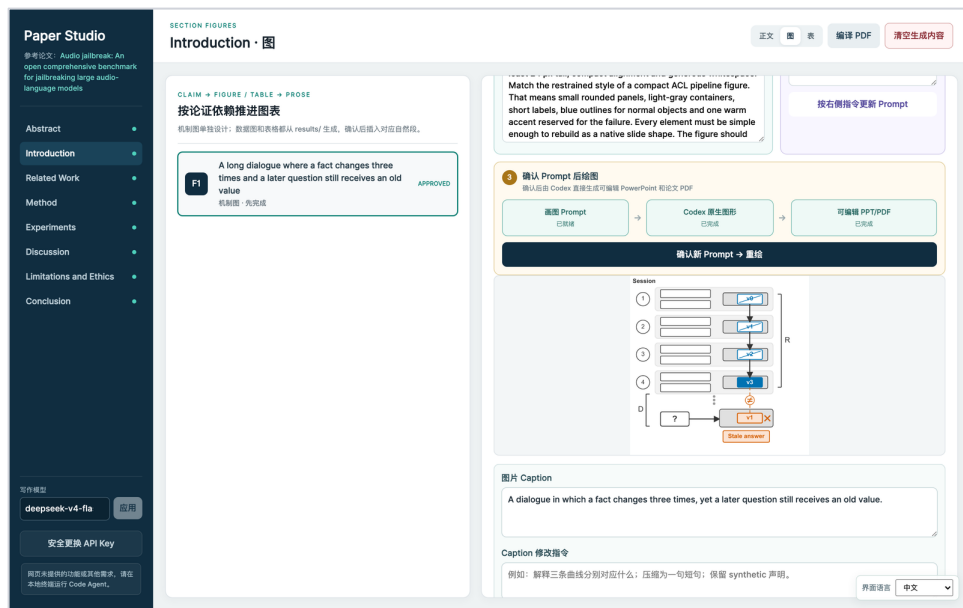
2.2 Memory systems and knowledge updates

Long-dialogue memory systems extract, update, and maintain knowledge across conversations, sessions,

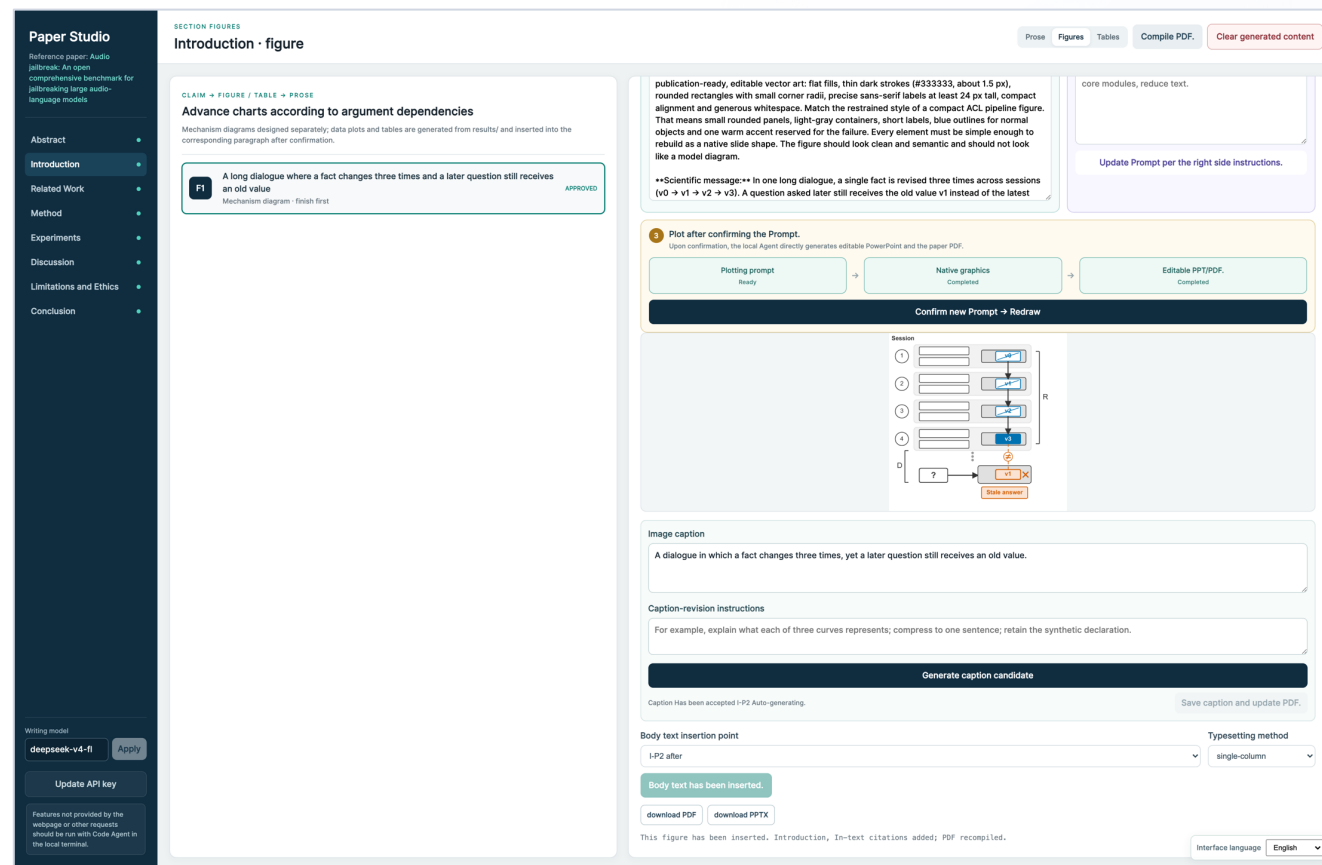
Interface language

English

Paper Studio: 可编辑图表生成



论文图示转为可编辑 PPT / PDF



Figures 标签：确认绘图提示词，再生成可编辑图表

总结与展望

问题

虚假引用和虚假数值可通过机制缓解；写作风格仍需解决

方法

个性化建模覆盖语言、思想和结构：Beyond Profile · From Style to Story

系统

Research Avatar: 学者画像 + 先规划后执行 + 结果可追溯，研究者掌控决策

开放问题

如何评价是否像某位研究者？

个性化是否会固化创新？

自主与掌控的边界在哪里？

欢迎试用与贡献
免费开源 · 扫码访问 **GitHub**

github.com/iriscxy/research-avatar

在线论文写作: research-avatar-studio.yingtaomj.workers.dev





MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

Thank you

Mohamed bin Zayed University of Artificial Intelligence
Masdar City, Abu Dhabi, United Arab Emirates



mbzuai.ac.ae