个体图灵测试：利用纵向个人数据进行基于大语言模型的模拟案例研究

Minghao Guo 罗格斯大学 美国新不伦瑞克，NJ
minghao.guo@rutgers.edu

Ziyi Ye* 复旦大学 中国上海
zyye@fudan.edu.cn

Wujiang Xu 罗格斯大学 美国新不伦瑞克，NJ
wujiang.xu@rutgers.edu

Xi Zhu 罗格斯大学 美国新不伦瑞克，NJ
xi.zhu@rutgers.edu

Wenyue Hua 微软 美国纽约州纽约市
wenyuehua@microsoft.com

Dimitris N. Metaxas 罗格斯大学 美国新泽西州新不伦瑞克
dnm@cs.rutgers.edu

arXiv:2603.01289v1 [cs.CL] 1 Mar 2026

摘要

大语言模型（LLM）展现出了令人瞩目的类人能力，但其复现特定个体的能力仍未得到充分探索。本文通过一项案例研究，探讨基于大语言模型的个体模拟，所用数据为一位志愿者提供的私人消息记录档案，时间跨度超过十年。基于这些消息数据，我们提出“个体图灵测试”，用于评估志愿者的熟人能否从多个候选回答中正确识别出最可能由该志愿者给出的回答。我们研究了常见的基于大语言模型的个体模拟方法，包括微调、检索增强生成（RAG）、基于记忆的方法，以及结合微调与 RAG 或记忆的混合方法。实证结果表明，当前基于大语言模型的模拟方法无法通过个体图灵测试；但当对目标个体的陌生人进行同样的测试时，这些方法的表现显著更好。此外，尽管微调能够提升模型在日常聊天中的模拟效果，使其更能体现个体的语言风格，但检索增强和基于记忆的方法在涉及个人观点和偏好的问题上表现更佳。这些发现揭示了在纵向上下文条件下，使用大语言模型进行个体模拟时，参数化方法与非参数化方法之间存在根本性的权衡。

CCS 概念

• **信息系统** → **检索模型与排序**; •**计算方法** → **自然语言生成**;语言资源.

关键词

大语言模型、个体模拟、用户建模

1 引言

大语言模型（LLM）近期展现出卓越的个体模拟能力，这得益于其强大的语言理解能力、隐式角色学习能力和推理能力 [15, 24, 25]。现有基于大语言模型的个体级模拟主要依赖参数化 方法（例如微调模型权重）或非参数化 方法（例如检索增强提示在推理时注入外部证据） [7, 13, 14, 18, 21, 23, 28, 29]。近期的基于智能体的框架进一步拓展了这些技术，以支持交互行为和长期记忆机制 [22, 27, 34]。

尽管取得了这些进展，现有的大多数基于大语言模型的个体模拟研究仍局限于较短的时间范围，通常以静态用户档案或简短的对话上下文为条件 [3–5]。值得注意的是，纵向个人数据编码了多个行为维度，包括日常互动中的风格模式（例如简洁性、表情符号和习惯性措辞），以及可从长期上下文中推断出的个人观点和偏好。另一方面，现有评估主要依赖自动化的语言相似度指标或其他大语言模型的判断，这些方法不足以评估机器生成的回复对于人类评判者而言，是否真的与目标个体的真实表达难以区分 [2, 6, 9, 10, 12, 19, 20, 32]。

为弥补这一空白，我们提出了个体图灵测试，旨在评估基于某个特定个体的纵向上下文生成的机器回复，是否能与该个体的真实表达难以区分。不同于侧重于一般类人性的传统图灵测试 [16, 26] ，我们的测试要求大语言模型学习目标个体的语言风格和个人观点。测试由该个体的熟人进行评估，并与面向陌生人的通用图灵测试进行比较（陌生人会获得该个体的基本档案）。通过这一评估，我们可以衡量大语言模型在评估者对该个体了解程度不同的情况下，生成难以与人类回复区分的内容的能力。

我们基于一位志愿者历时十余年的私人消息记录纵向数据集开展评估，并比较了四种常见的个体模拟方法：微调、检索增强生成（RAG）、基于记忆的方法，以及融合其中两种方法的混合方法。表 1 展示了一个定性示例，其中真实回复的平均排名低于其他回复。这表明，当前的大语言模型无法从熟人的视角令人信服地模拟该个体。然而，通用图灵测试的证据表明，在以该个体的陌生人为评估对象时，大语言模型能够通过测试。此外，我们发现了风格与内容的权衡：参数化微调能够捕捉语言风格，而非参数化检索机制则能确保观点的一致性。此外，我们发现，模拟保真度会受到历史背景时间范围的

*通讯作者。

表 1：案例研究：对“你更向往大城市的机会，还是小城市较慢的生活节奏？**”的回复进行定性比较。平均排名表示人类评判者给出的平均名次（越低越好）。**

方法	回复内容	平均排名
标准答案	大城市。你能遇见各种各样的人，接触到更多新鲜事物。	**2.00**
A-Mem+LoRA	小城市的节奏对我来说太慢了，我实在没法那样生活。我宁愿出去，到别的地方工作。	**2.33**
检索增强生成+LoRA	大概是因为机会吧。大城市能提供更多可能性。	3.17
LoRA	啊，这个问题有点难回答。两者都各有魅力！你呢？	3.33
检索增强生成+Base	我确实很羡慕大城市有那么多机会。你喜欢小城镇的宁静吗？当然，大城市的竞争也很激烈。你呢？	3.50
A-Mem+Base	我更倾向于大城市的机会——那里确实更多。不过有时我也想放慢脚步，享受小镇的宁静。	3.67

影响，这凸显了纵向数据的必要性。

总之，这项工作有三点贡献。(1) 我们提出了**个体图灵测试**，这是一种用于评估个体模拟的协议，揭示了一般类人性与身份层面的真实性之间的差距。(2) 我们比较了**参数化**、**非参数化**和**混合**个性化策略，揭示了风格对齐与观点落地之间的权衡。(3) 我们证明**时间范围**会影响性能，利用近期历史数据带来的提升在超过某个阈值后趋于饱和。

2 方法论

2.1 任务定义

个体图灵测试旨在真实的对话情境下，生成与特定目标个体的响应难以区分的回答。给定目标个体 u 和一组对话提示 Q，目标个体针对每个 $q \in$ Q 生成一个真实响应 $r^{\text{gt}}(q)$。模拟方法 m 根据相同的提示 q 生成相应的响应 $r_{(m)}(q)$。随后，我们通过严格的人工判断和自动化评估指标，评估模拟响应能否忠实反映目标个体的真实行为。

2.2 数据集构建

为实现个体特异性和长期评估，我们基于一位志愿者贡献的纵向私人消息档案开展了一项研究。研究通过一家匿名机构获得了伦理审批。该数据集涵盖约十年的日常对话，涉及广泛的日常与个人互动场景。所有数据均经过匿名化处理，并移除了敏感个人信息。我们采用两阶段时间窗口启发式方法构建对话实例：同一说话者连续发送的消息在 2 分钟内合并，用户与助手的轮次则使用 5 分钟窗口进行配对。为进一步提升数据质量，我们结合基于重叠的启发式方法（子串／相等性检查，以及针对简短回复的 > 0.8 字符覆盖重叠）和轻量级黑名单，剔除重复且信息量低的回复（例如确认语和媒体占位符）。该数据集共包含 12,151 段对话、72,652 条消息和 1,157,842 个训练词元。值得注意的是，该数据集的规模远大于以往研究所用的数据集，因此能够更全面地评估个体模拟的泛化能力和长期一致性。

2.3 评估协议

我们采用一种混合式评估框架，结合**人工判断**与**自动语言相似度指标**，针对两种具有代表性的提示类型——**日常对话**和**个人观点**——进行评估，以全面衡量个体模拟的整体保真度及其内在行为机制。

人工判断：个体图灵测试。 我们将个体图灵测试 形式化为一项基于排名的识别任务。对于每个提示 $q \in$ Q，评估者会看到一个经随机打乱的候选池，其中包含 (i) 来自目标个体的真实参考回复 $r^{\text{gt}}(q)$，以及 (ii) 由不同个体模拟方法生成的模拟响应 $\{r_{(m)}(q)\}_{m \in \mathcal{M}}$。随后，评估者根据各回复与目标个体相符的合理程度对所有回复进行排序（排名越低越合理）。评估者包括该个体的 7 位熟人，他们熟悉该个体且友谊至少持续 3 年，但明确排除该个体的直系亲属。

此外，我们在相同的人工判断协议下开展一项辅助性通用图灵测试，以检验通用能力。在通用图灵测试中，回复由 5 位陌生人评估，这些陌生人会获得目标个体的基本档案，包括工作经历、教育背景和人口统计信息。

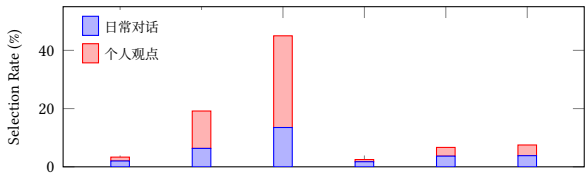
基于排名数据，我们计算选择率（SR），定义为某种方法排名第一的次数占比，用以反映最有把握的识别结果。对于每种方法 m，我们计算其 选择率

$$\mathrm{SR}(m) = \frac{C_m}{\sum_{m' \in \mathcal{M} \cup \{\text{gt}\}} C_{m'}},$$

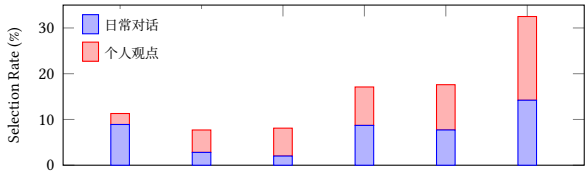
其中，C_m 表示人类评估者选择方法 m 生成的回答的次数。在不同方法中，选择率低于标准答案表明，在人工判断下仍存在明显的可区分性差距；而选择率越接近，则表明个体层面的真实性越强。

自动语言相似度指标。除了人工判断外，我们还计算了 BLEU-1/2、ROUGE-L、精确率、召回率和 Distinct-2 等自动指标，以反映词汇准确性、结构、内容覆盖率和词汇多样性。

提示类型。 我们考察了两种具有代表性的提示类型，它们在时间和语义侧重点上有所不同，每类抽取 30 条提示。**日常对话** 涵盖日常及情境性互动，主要反映即时上下文和时效性，与以往研究中常见的日常闲聊和开放领域对话设置相一致（例如， [11, 33]），例如“你最近看到了什么新闻？”。相比之下，**个人观点** 探究更长期的偏好或判断倾向，这些倾向预计会随时间保持一致，符合个性化和用户建模研究中以观点和偏好为导向的提示（例如， [1, 31, 33]），例如“阅读新闻时，你更在意事实还是观点？”。



(a) 通用图灵测试（陌生人， $N = 5$）。



(b) 个体图灵测试 (熟人， $N = 7$)。

图1：通用图灵测试与个体图灵测试的比较。我们报告了两种评估设置下各方法的选择率：(a) 通用图灵测试，**由**陌生人**评判回复；(b)** 个体图灵测试，**由目标个体的熟人评判回复。结果按提示类型堆叠，区分**日常对话**提示和**个人观点**提示**。

3 实验结果与讨论

我们从三个角度评估基于大语言模型的个体模拟。首先，我们比较个体图灵测试（由熟人进行判断）和一般图灵测试（由陌生人进行判断）下的表现，并将后者作为衡量大语言模型总体类人性的基线。其次，我们考察不同模拟策略在个体图灵测试下的表现。最后，我们分析个体消息历史记录的时间范围对大语言模型表现的影响。

模拟方法。 为研究不同个体模拟方法对大语言模型性能的影响，我们系统比较了参数化方法与非参数化方法。我们采用 LoRA-7B [8] 作为参数化设置，因为研究发现，在纵向数据上进行完整微调会导致性能显著下降。对于非参数化建模，我们同时考察了基于检索的 RAG [17] 和基于记忆的 A-Mem [30]。此外，我们还评估了将 LoRA 与 RAG 或 A-Mem 结合的混合变体。

实现细节。 我们选择 Qwen2.5-7B 作为所有方法的模拟骨干大语言模型。对于 LoRA，我们在 $2 \times$ A800 GPU 上对模型进行微调，秩为 $r = 4$，dropout 为 0.3，目标投影模块为（q_proj、v_proj）。模型训练 2 个 epoch，学习率为 1×10^{-4}，并采用余弦调度。对于 RAG/A-Mem/Hybrid 推理，我们使用 BGE-M3 嵌入检索 top-$k = 5$ 条记忆，并采用余弦相似度过滤（最小_余弦相似度 $= 0.35$）和去重（去重_余弦相似度 $= 0.92$）。推理时采用受控解码，温度为 0.85，重复惩罚为 1.2，最大 token 数为 80。

统计分析方案。 为严格评估性能差异，我们根据排名差异进行了事后两两比较。除非另有说明，显著性均通过单侧二项检验确定。

3.1 图灵测试：基于大语言模型的个体模拟能力

我们首先对比分析大语言模型在通用图灵测试 和个体图灵测试中的表现，如图1所示。在通用图灵测试中，真实回答并未取得最高选择率。相反，基于大语言模型的模拟方法（例如 A-Mem+Base）取得了相当甚至更高的选择率，这表明在陌生人评估时，基于大语言模型的模拟能够生成看似符合人类表达习惯的回答。例如，事后两两比较显示，A-Mem+Base 的回答显著优于真实参考答案，达到 $p = 4.79 \times 10^{-8}$。

相比之下，在个体图灵测试中，真实回答的选择率最高。事后两两比较证实，真实回答显著优于表现最佳的模拟方法 A-Mem+LoRA（$p = 6.21 \times 10^{-3}$）。这表明，从亲密熟人的视角来看，当前大语言模型仍难以忠实地复现目标个体的行为。这两种评估设置之间的差异表明，具备一般的类人性并不意味着具备身份特异性保真度。

3.2 不同模拟方法的比较

性能层级。 总体而言，如图 1b 所示，混合方法表现最佳，其次是仅LoRA方法，而仅检索和仅记忆方法表现最差。成对比较证实了这一层级：RAG+LoRA 显著优于 RAG+Base（胜率 $= 0.694$，$p = 2.88 \times 10^{-11}$），仅LoRA方法也显著优于 RAG+Base（胜率 $= 0.608$，$p = 5.59 \times 10^{-4}$），而 RAG+LoRA 的表现又进一步超过仅LoRA方法（胜率 $= 0.614$，$p = 5.72 \times 10^{-5}$）。综上，这些结果确立了清晰且一致的性能层级：混合 > 仅参数化 > 仅非参数化。

表 2：选定的自动化指标（诊断性）。混合方法（LoRA + 检索/记忆）始终优于仅使用 LoRA 和仅使用检索/记忆的基线方法。

方法	BLEU-1 ↑	BLEU-2 ↑	ROUGE-L ↑	精确率. ↑	召回率. ↑	Dist-2 ↑
仅使用 LoRA	0.1578	0.0501	**0.1496**	0.1565	0.2128	0.7171
检索增强生成+基线	0.1011	0.0264	0.1089	0.1005	0.2949	0.6407
A-Mem+基线	0.1113	0.0309	0.1228	0.1109	**0.3613**	0.6433
检索增强生成+LoRA	0.1837	0.0735	0.1456	**0.1776**	0.1714	0.7680
A-Mem+LoRA	**0.1983**	**0.1175**	0.1457	0.1649	0.1586	**0.7881**

不同提示类型之间的机制差异。 为理解这一层级，我们分析了不同提示类别下的表现。参数化适配（LoRA）主要有助于以风格为主导的日常互动，在这类互动中，匹配对话语气、简洁性和非正式措辞至关重要。相比之下，非参数化检索和记忆机制（检索增强生成和 A-Mem）对涉及个人观点和偏好的提示贡献更大，因为与有历史依据的观点保持一致至关重要。换言之，参数化方法主要决定风格，而非参数化记忆则确保内容，混合模型结合两者优势，从而在两种提示类型上取得更出色的表现。

我们通过一个具体示例来说明这种差异。当被问到“你通常喜欢什么类型的电影或电视节目？”时，仅使用 LoRA 的模型会给出简短且风格契合的回复，例如“我喜欢悬疑片和科幻片。你呢？”。相比之下，RAG+Base 会检索具体的历史参照（例如“我主要喜欢动作片和奇幻片——我特别喜欢像《指环王》这样的作品。”），但措辞听起来可能不够个性化。混合式 RAG+LoRA 兼具风格对齐与有依据的细节：“我主要喜欢科幻片，比如《星际穿越》。你呢？”。这一质性对比与图 1b 中观察到的量化趋势相呼应。

与自动化指标的对齐。 这些基于人工判断的发现也体现在自动化指标中（表 2）。BLEU-1 衡量单元词重叠，因此反映了与目标个体之间的词元级词汇对齐；BLEU-2 和 ROUGE-L 则捕捉局部及序列级结构一致性。精确率通过衡量生成词元中出现在参考回复里的比例，对过度生成或分布外生成进行惩罚；召回率则鼓励更广泛地覆盖基于历史的内容。Distinct-2 量化二元词组多样性，能够反映模式崩溃的敏感程度。从这一角度看，参数化适配往往能提升 BLEU-1、精确率和 Distinct-2，表明其风格约束更强，且多样性更可控。相比之下，检索和记忆机制对召回率及词汇覆盖率指标的提升更为显著，体现了它们在以历史内容为依据生成回复方面的作用。混合模型兼顾这两种效果，取得了整体表现最优的指标表现。

3.3 消息历史记录时间范围的影响

我们进一步分析个人消息历史记录的时间范围如何影响大语言模型的表现。我们采用性能最强的混合配置（A-Mem + LoRA），将消息历史记录的时间范围从最近

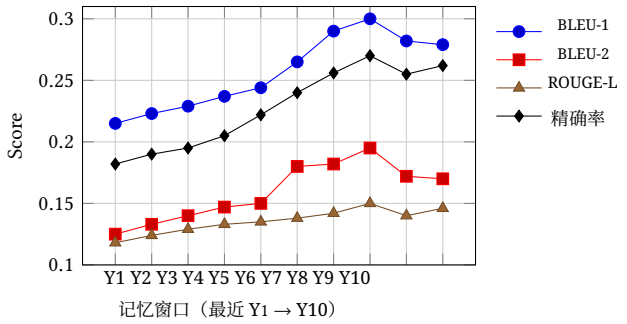


图 2：扩展近期记忆对混合模拟的影响（A-Mem + LoRA）。 Y_i 表示使用最近的 i 年对话历史（例如， Y_1 = 仅使用最近一年； Y_{10}=最近十年）。

一年扩展至最近十年，并使用自动语言相似度指标评估大语言模型的表现。图 2 显示，随着纳入更多近期年份，模型表现持续提升；对话历史约扩展至八年时，提升幅度仍然稳定。BLEU-1 和精确率单调递增，表明词元级一致性和风格对齐有所增强；BLEU-2 和 ROUGE-L 也逐步提升，说明词汇锚定得到加强。

然而，当时间跨度延长至约八年以上时，表现趋于平稳并略有下降。纳入更久远的历史可能会引入过时或时间相关性较低的模式，从而稀释近期行为背景中的信号。这些结果表明，有效的个性化不仅取决于纳入的记忆量，也取决于时间相关性。扩展近期历史记录可在达到某个阈值前提升依据充分程度；超过该阈值后，进一步扩大记忆覆盖范围带来的表现提升微乎其微。

4 结论

我们通过通用图灵测试和个体图灵测试，研究了大语言模型在个体模拟中的表现。尽管混合方法（将参数适配与检索或记忆相结合）始终优于单一组件方法，但所有模型的表现仍远低于人类，显现出持续存在的真实性差距。值得注意的是，两种评估设置之间的差异表明，在熟人判断中，具备一般类人性并不意味着能够实现身份层面的保真度。

我们的研究结果表明，个体模拟中存在结构性分工：参数化适配主要捕捉风格规律，而非参数化记忆机制则保持观点一致性与基于历史的内容。此外，有效的个性化不仅取决于记忆覆盖范围，还取决于时间相关性，因为扩展近期历史对性能的提升仅持续到饱和阈值为止。

这些结果表明，忠实的个体模拟需要协调风格、内容和时间对齐。未来的研究不应仅着眼于提升原始生成质量，还应进一步探究如何在不同的对话情境下，动态平衡参数化适配与记忆的依据。未来

的研究还可以探索更加系统的记忆选择策略和考虑近期性的架构，以进一步缩小真实性差距，更好地捕捉身份层面的行为信号。

参考文献

[1] 世界价值观调查：第六轮——国家合并数据文件版本。www.worldvaluessurvey.org/WVSDocumentationWV6.jsp[2] Elizabeth Clark、Tal August、Sofia Serrano、Nikita Haduong、Suchin Gururangan 和 Noah A. Smith。2021年。《并非所有“人类”之物都是黄金：评估对生成文本的人工评估》。载于第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），Chengqing Zong、Fei Xia、Wenjie Li 和 Roberto Navigli（编）。计算语言学协会，线上，7282–7296。doi:10.18653/v1/2021.acl-long.565[3] Bangde Du、Minghao Guo、Songming He、Ziyi Ye、Xi Zhu、Weihang Su、Shuqi Zhu、Yujia Zhou、Yongfeng Zhang、Qingyao Ai 和 Yiqun Liu。2025年。《Twin-Voice：通过大语言模型角色模拟构建数字孪生多维基准》。arXiv: 2510 25536 [ ] 2510 25536 。[4] cs.CL https://arxiv.org/abs/ 。Bangde Du、Ziyi Ye、Monika Jankowska、Zhijing Wu、Qingyao Ai 和 Yiqun Liu。2025年。《理解短视频浏览中观点极化的影响》。arXiv: 2403 04184 [ ] 2403 04184 。cs.SI https://arxiv.org/abs/ 。[5] Bangde Du、Ziyi Ye、Zhijing Wu、Monika A. Jankowska、Shuqi Zhu、Qingyao Ai、Yujia Zhou 和 Yiqun Liu。2025年。《SimVBG：通过背景故事生成模拟个体价值观》。载于*2025*年自然语言处理实证方法会议论文集，Christos Christodoulopoulos、Tanmoy Chakraborty、Carolyn Rose 和 Violet Peng（编）。计算语言学协会，中国苏州，13093 13122 10 18653 1 2025 662 – 。doi: 。/v / .[6] emnlp-main。Sebastian Gehrmann、Tosin Adewumi、Karmanya Aggarwal、Pawan Sasanka Am- manamanchi、Anuoluwapo Aremu、Antoine Bosselut、Khyathi Raghavi Chandu、Miruna Clinciu、Dipanjan Das、Kaustubh Dhole 等。2021年。《GEM基准测试：自然语言生成、评估及其指标》。载于第一届自然语言生成、评估与指标研讨会论文集（*GEM 2021*）。96–120。[7] Minghao Guo、Qingcheng Zeng、Xujiang Zhao、Yanchi Liu、Wenchao Yu、Meng- nan Du、Haifeng Chen 和 Wei Cheng。2026年。《DeepSieve：通过大语言模型知识路由器进行信息筛选》。arXiv:2507.22050 [cs.CL] https://arxiv.org/ abs/2507.22050[8] Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。2021年。《LoRA：大语言模型的低秩适配》。arXiv:2106.09685 [cs.CL] https://arxiv.org/abs/2106.09685[9] Wenyue Hua、Xianjun Yang、Mingyu Jin、Zelong Li、Wei Cheng、Ruixiang Tang 和 Yongfeng Zhang。2024年。《Trustagent：迈向安全、可信的大语言模型智能体》。载于计算语言学协会论文集：*EMNLP 2024*。10000–10016。[10] Wenyue Hua、Kaijie Zhu、Lingyao Li、Lizhou Fan、Mingyu Jin、Shuhang Lin、Haochen Xue、Zelong Li、JinDong Wang 和 Yongfeng Zhang。2025年。《解构逻辑：上下文在大语言模型推理能力中的作用》。载于计算语言学协会论文集：*ACL 2025*。19219–19242。[11] Pegah Jandaghi、Xianghai Sheng、Xinyi Bai、Jay Pujara 和 Hakim Sidahmed。2024年。《利用大语言模型生成忠实的基于角色的对话数据集》。载于第六届对话式人工智能自然语言处理研讨会论文集（*NLP4ConvAI 2024*），Elnaz Nouri、Abhinav Rastogi、Georgios Spithourakis、Bing Liu、Yun-Nung Chen、Yu Li、Alon Albalak、Hiromi Wakaki 和 Alexandros Papangelis（编）。计算语言学协会，泰国曼谷，114–139。https://aclanthology.org/2024.nlp4convai-1.8/[12] Jianchao Ji、Yutong Chen、Mingyu Jin、Wujiang Xu、Wenyue Hua 和 Yongfeng Zhang。2025年。《Moralbench：大语言模型的道德评估》。*ACM SIGKDD*探索通讯 27，1（2025），62–71。[13] Mingyu Jin、Weidi Luo、Sitao Cheng、Xinyi Wang、Wenyue Hua、Ruixiang Tang、William Yang Wang 和 Yongfeng Zhang。2025年。《解构大语言模型中的记忆与推理能力》。载于第*63*届计算语言学协会年会论文集（第*1*卷：长篇论文）。1681–1701。[14] Mingyu Jin、Qinkai Yu、Jingyuan Huang、Qingcheng Zeng、Zhenting Wang、Wenyue Hua、Haiyan Zhao、Kai Mei、Yanda Meng、Kaize Ding 等。2025年。《探索概念深度：大语言模型如何在不同层级获取知识和概念？》。载于第*31*届国际计算语言学会议论文集。558–573。[15] Mingyu Jin、Qinkai Yu、Dong Shu、Haiyan Zhao、Wenyue Hua、Yanda Meng、Yongfeng Zhang 和 Mengnan Du。2024年。《推理步骤长度对大语言模型的影响》。载于计算语言学协会论文集：*ACL 2024*。1830–1842。[16] Douwe Kiela、Max Bartolo、Yixin Nie、Divyansh Kaushik、Atticus Geiger、Zhengx- uan Wu、Bertie Vidgen、Grusha Prasad、Amanpreet Singh、Pratik Ringshia、Zhiyi Ma、Tristan Thrush、Sebastian Riedel、Zeerak Waseem、Pontus Stenetorp、Robin Jia、Mohit Bansal、Christopher Potts 和 Adina Williams。2021年。《Dynabench：重新思考自然语言处理基准测试》。arXiv:2104.14337 [cs.CL] https: // arxiv.org/abs/2104.14337[17] Patrick Lewis、Ethan Perez、Aleksandra Piktus、Fabio Petroni、Vladimir Karpukhin、Naman Goyal、Heinrich Küttler、Mike Lewis、Wen tau Yih、Tim Rocktäschel、Sebastian Riedel 和 Douwe Kiela。2021年。《面向知识密集型自然语言处理任务的检索增强生成》。arXiv:2005.11401 [cs.CL] https://arxiv.org/abs/ 2005 11401 。Danrui Li、Sen Zhang、Samuel S. Sohn、Kaidong Hu、Muhammad Usman 和 Mubbasir Kapadia。2025年。《Cardiverse：利用大语言模型进行新颖卡牌游戏原型设计》。载于*2025*年自然语言处理实证方法会议论文集，Christos Christodoulopoulos、Tanmoy Chakraborty、Carolyn Rose 和 Violet Peng（编）。计算语言学协会，中国苏州，29735–29762。doi:10.18653/v1/2025.emnlp-main.1511[19] Haitao Li、Qian Dong、Junjie Chen、Huixue Su、Yujia Zhou、Qingyao Ai、Ziyi Ye 和 Yiqun Liu。2024年。《大语言模型评审：基于大语言模型的评估方法综合综述》。arXiv: 2412 05579 [ ] 2412 05579 。cs.CL [20] https://arxiv.org/abs/ 。Shuhang Lin、Wenyue Hua、Lingyao Li、Che-Jui Chang、Lizhou Fan、Jianchao Ji、Hang Hua、Mingyu Jin、Jiebo Luo 和 Yongfeng Zhang。2024年。《Battleagent：对历史战役进行多模态动态模拟，以补充历史分析》。载于*2024*年自然语言处理实证方法会议论文集：系统演示。172–181。[21] Shuhang Lin、Zhencan Peng、Lingyao Li、Xiao Lin、Xi Zhu 和 Yongfeng Zhang。2025年。《智能体检索增强生成系统的缓存机制》。arXiv:2511.02919 [cs.CL] https://arxiv.org/abs/ 2511 02919 。Weidi Luo、Qiming Zhang、Tianyu Lu、Xiaogeng Liu、Bin Hu、Hung-Chun Chiu、Siyuan Ma、Yizhe Zhang、Xusheng Xiao、Yinzhi Cao 等。2025年。《代码智能体可以成为端到端系统黑客：计算机操作智能体现实世界威胁的基准测试》。*arXiv*预印本 *arXiv:2510.06607*（2025）。[23] Kai Mei、Jiang Guo、Shuaichen Chang、Mingwen Dong、Dongkyu Lee、Xing Niu 和 Jiarong Jiang。2025年。《R-wom：面向计算机操作智能体的检索增强世界模型》。*arXiv*预印本 *arXiv:2510.11892*（2025）。[24] Joon Sung Park、Joseph C. O’Brien、Carrie J. Cai、Meredith Ringel Morris、Percy Liang 和 Michael S. Bernstein。2023年。《生成式智能体：人类行为的交互式模拟体》。arXiv: 2304 03442 [ ] 2304 03442 。cs.HC [25] https://arxiv.org/abs/ 。Yunfan Shao、Linyang Li、Junqi Dai 和 Xipeng Qiu。《Character-LLM：用于角色扮演的可训练智能体》。arXiv:2310.10158 [cs.CL] https://arxiv.org/ abs/2310.10158[26] 艾伦·图灵。2007年。《计算机器与智能》。载于解析图灵测试：寻找思考型计算机过程中的哲学与方法论问题。施普林格，23–65。[27] Qingyue Wang、Yanhe Fu、Yanan Cao、Shuai Wang、Zhiliang Tian 和 Liang Ding。2025年。《递归式摘要使大语言模型具备长期对话记忆》。arXiv:2308.15022 [cs.CL] https://arxiv.org/abs/2308.15022[28] Ziqi Wang、Xi Zhu、Shuhang Lin、Haochen Xue、Minghao Guo 和 Yongfeng Zhang。2026年。《RAGRouter-Bench：自适应检索增强生成路由的数据集与基准测试》。*arXiv*预印本 *arXiv:2602.00296*（2026）。[29] Rubin Wei、Jiaqi Cao、Jiarui Wang、Jushi Kai、Qipeng Guo、Bowen Zhou 和 Zhouhan Lin。2025年。《MLP记忆：面向大语言模型的检索器预训练记忆》。arXiv: 2508 01832 [ ] 2508 01832 。cs.CL [30] https://arxiv.org/abs/ 。Wujiang Xu、Zujie Liang、Kai Mei、Hang Gao、Juntao Tan 和 Yongfeng Zhang。2025 年。《 A-Mem：大语言模型智能体的智能体记忆》。*arXiv* 预印本 *arXiv:2502.12110*（2025）。[31] Ziyi Ye、Xiaohui Xie、Yiqun Liu、Zhihong Wang、Xuesong Chen、Min Zhang 和 Shaoping Ma。2022年。《利用脑信号更好地理解人类阅读理解》。载于*2022*年*ACM*网络会议论文集。380–391。[32] Qingcheng Zeng、Mingyu Jin、Qinkai Yu、Zhenting Wang、Wenyue Hua、Guangyan Sun、Yanda Meng、Shiqing Ma、Qifan Wang、Felix Juefei-Xu 等。2025年。《选择题的不确定性量化只需单令牌深度》。载于第*34*届*ACM*信息与知识管理国际会议论文集。5474–5478。[33] Saizheng Zhang、Emily Dinan、Jack Urbanek、Arthur Szlam、Douwe Kiela 和 Jason Weston。2018年。《对话智能体个性化：我有只狗，你也养宠物吗？》。arXiv: 1801 07243 [ ] 1801 07243 。cs.AI [34] https://arxiv.org/abs/ 。Yinhe Zheng、Guanyi Chen、Minlie Huang、Song Liu 和 Xuan Zhu。2020年。《具有多样化特征的个性化对话生成》。arXiv:1901.09672 [cs.CL] https://arxiv.org/abs/1901.09672