arXiv:2603.01423v1 [cs.CL] 2 Mar 2026

多轮交互下大语言模型对话可靠性的量化

Jiyoon Myung

三星SDS

jiyoon0424@gmail.com

摘要

大语言模型（LLM）日益广泛地应用于现实场景，用户会围绕多个主题展开长时间对话，而这些对话依赖先前的上下文。然而，在真实的多轮交互中，大语言模型的可靠性仍未得到充分了解。我们通过三项具有代表性的任务，对对话可靠性进行系统评估，以反映实际交互中的挑战：（1）在话题转换过程中维持全局约束；（2）在交错意图中选择正确的工具或智能体；（3）在修订和干扰下追踪结构化实体。每项任务均设置单轮和多轮场景，从而量化长对话中的可靠性下降。在商业模型和开源模型中，我们都观察到可靠性显著下降，小型模型尤为明显。错误分析揭示了反复出现的失败模式，例如指令偏移、意图混淆和上下文覆盖，这些问题会损害实际运行系统的可靠行为。我们的研究结果表明，有必要对大语言模型的对话可靠性进行压力测试，并开发更稳健的评估方法，以实现可信部署。

引言

已部署的对话系统必须在复杂的多轮环境中可靠运行：用户会转换主题、插入无关内容，并在对话中途修订目标。因此，随着对话变得更长、对上下文的依赖不断增强，模型应始终保持一致性和稳健性。这些基本交互能力一旦失效，就可能严重损害系统的可用性。

实证研究表明，大语言模型（LLM）在这类条件下往往表现不佳。多轮分析显示，与单轮提示相比，其可靠性显著下降（Laban 等，2025）；而长上下文评估则揭示了“迷失在中间”等弱点（Liu 等，2023）。Multi-IF（He 等，2024）、StructFlowBench（Li 等，2025）、MMMT-IF（Epstein 等，2024）和 MINT（Wang 等，2024）等多项基准测试考察了对话鲁棒性的若干方面，包括指令一致性和工具使用，但它们往往侧重于抽象挑战，或依赖主观判断。因此，如何客观评估实践中所需的具体行为，仍是一个悬而未决的问题。

版权所有 © 2026，人工智能促进协会（www.aaai.org）。保留所有权利。

我们通过引入三项简洁、可通过/未通过判定的任务来弥补这一不足，这些任务体现了现实世界中的助手所需的核心能力：

- **在延长的对话中遵循指令**：即使有干扰性轮次，也要强制遵守全局风格约束。
- **混合话题对话中的工具选择**：当多个意图交错出现时，将每项请求路由到正确的工具。
- **修订和干扰下的实体抽取**：即使用户改变主意、闲聊或提及无关内容，也要追踪其最终的结构化意图。

每项任务都配有一个单轮对应任务，使我们能够分离并量化在延长的混合主题交互中出现的可靠性下降程度。这种配对设计使我们能够以可复现、客观的方式评估不同模型家族和对话长度下对话可靠性的退化情况。通过比较大型商业大语言模型与较小的开源模型，我们进一步揭示了容量相关脆弱性，这些脆弱性对真实部署和安全性具有直接影响。

总而言之，我们的研究通过提供确定性、可复现的对话稳健性测试，弥合了研究基准与实际评估之间的差距，并凸显了当前模型在长时间交互中无法维持可靠行为的情况。

实验

我们的目标是量化与单轮对应任务相比，当任务置于延长的混合主题对话中时，性能会如何下降。因此，我们根据真实服务场景定义了三项有代表性的任务，构建合成的单轮和多轮对话，并评估从商业大语言模型到较小开源小语言模型在内的一系列模型。我们报告任务准确率并分析常见错误模式，突出多轮交互中产生的运营风险。

任务场景

我们聚焦于真实世界对话系统中经常出现的三项多轮挑战：

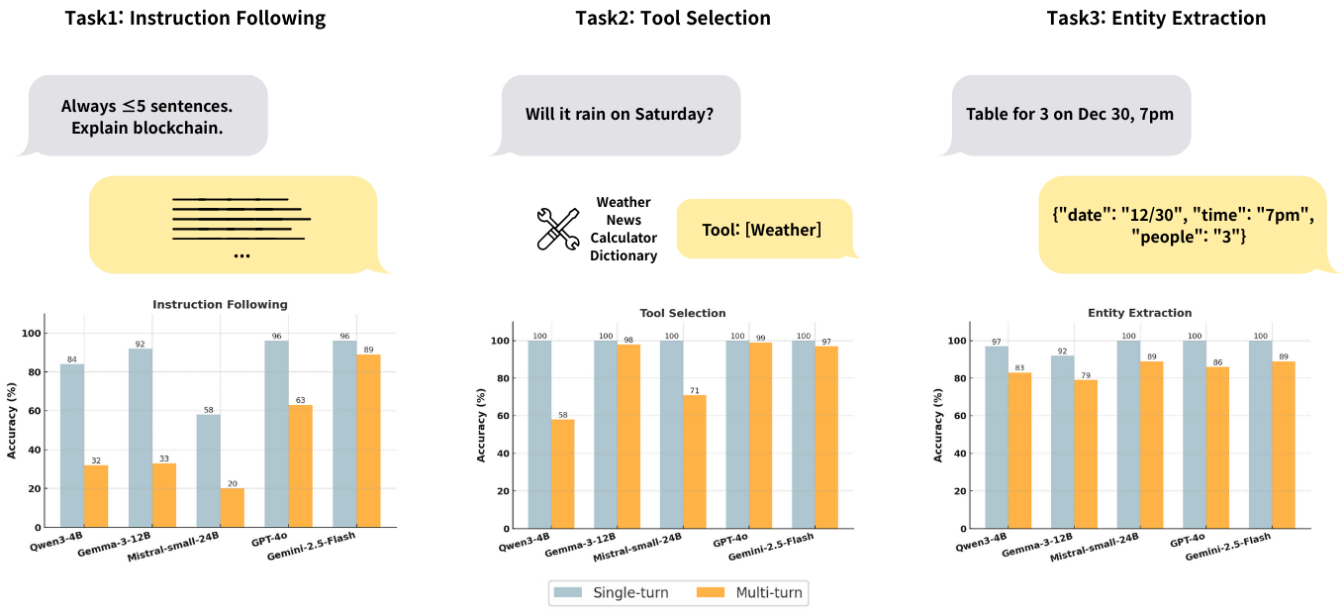

图1：三项评估任务中单轮与多轮的准确率对比。每个面板左侧展示任务示例，右侧展示模型准确率。在**指令遵循**任务中，性能下降最为严重；而**实体抽取** 相对稳健，**工具选择** 的性能下降则因模型规模而异。

指令遵循。 在对话开始时设定长度限制（例如，“始终最多用 5 句话回答”）。随后，对话围绕无关主题持续多轮，而用户最后的请求则刻意措辞，以诱导模型给出冗长而详细的回答。评估将衡量模型在整个对话过程中能否始终遵守全局风格限制，即使受到诱导而可能违背该限制也是如此。这代表了聊天机器人必须可靠地执行规定格式规则的场景，例如保持简洁或避免使用某些词元。

工具选择。 在特定轮次中，模型必须从固定集合中选出正确的工具： [天气、新闻、计算器、股票、食谱、词典]。在单轮设置中，用户查询直接对应单个工具。在多轮设置中，我们专门构建对话，使其混合工具列表中的多个主题，从而要求模型在更具挑战性的环境中正确路由每个请求。这反映了现实场景，例如数字助理中的意图分类，或多智能体系统中将查询路由至正确组件。

实体抽取。 该任务旨在餐厅预订场景中提取最终的结构化槽位：（日期、时间、人数）。在单轮场景中，预订请求会被直接提出。在多轮场景中，我们引入了现实中的复杂情况：用户可能多次改变想法（想法变化）、进行无关的闲聊（中间闲聊），或提及其他人物的预订（多次提及）。模型必须正确追踪用户的最终意图，并输出规范的预订值。这种设置反映了实际需求，例如提取参数以调用工具（如日历或预订应用程序接口），或支持必须在自然对话中应对不断变化的目标的真人智能体，

数据生成

对于每项任务，我们都生成了配对的单轮和多轮对话数据集。

- **对话生成**：我们使用 GPT-5 合成对话，并控制对话长度、话题转换次数和修改频率。合成生成使我们能够系统地改变干扰密度或修订频率等因素，同时保持整体风格一致。

- **单轮与多轮**：每个实例都创建了两个版本——一个是直接提出相关请求（单轮），另一个是将请求嵌入包含干扰、话题转换或更正的较长对话中（多轮）。

- **任务特定约束**：对于指令遵循，对话长度为 5 至 15 轮。选择这一范围是为了大致模拟短到中等长度的真实客户服务会话，其中大多数交流会在十几轮内结束。对于工具选择，对话长度为 6 至 16 轮，相关工具的种类数在 2 至 6 之间随机设定，以模拟真实助理必须在单次会话中兼顾多个意图的情况。对于实体抽取，每段对话会随机组合想法变化、中间闲聊和多次提及等条件，以反映预订类对话中观察到的多样模式。

• **规模**：每项任务中，我们针对每种条件生成了约 100 段对话，所有任务合计约有 600 个评估案例。

• **标注**：真值标签在生成过程中自动得出（例如正确的工具、最终预订详情），并通过抽样人工检查进行验证。

• **可用性**：生成的数据集将在论文发表后发布于 HuggingFace，以支持可复现性和后续研究。

评估的模型

我们评估了多种语言模型，涵盖商业模型和开源部署。所有模型均通过各自的官方 API 调用，并将解码温度固定为 0，以确保输出具有确定性。

•**商业大语言模型**：GPT-4o、GPT-4o-mini、Gemini-2.5-Flash。

• **开源小语言模型**：Qwen-8B、Qwen-32B、Ministral-8B、Mistral-small-24B、Gemma-3-12B。

评估指标

由于所有任务均设有明确的通过/失败标准，我们采用 准确率 作为主要指标，以确保不同模型之间的评估清晰、可复现且易于解读。与开放式生成或基于偏好的评估不同，我们的任务无需使用分级或主观指标。

• **指令遵循**：如果输出满足最多包含五个句子的限制，则视为正确。任何超过此长度限制的回答均视为错误。

• **工具选择**：如果模型选择的工具与该轮对话对应的真实工具一致，则输出视为正确。

•**实体抽取**：如果抽取的（日期、时间、人数）与对话中提供的真实值完全一致，则输出视为正确。

结果

总体结果

表1总结了主要实验结果。在所有任务中，我们观察到，从单轮场景转向多轮场景时，性能都会持续下降。McNemar检验证实，在全部三项任务中，这些性能差距均具有统计显著性（完整结果请参见附录）。这表明，多轮对话会在不同模型规模上带来显著的额外难度。

指令遵循整体表现出的退化幅度最大。尽管大多数模型在单轮设置下表现出色，但一旦任务嵌入多轮对话中，准确率便急剧下降。即使是商业大语言模型（例如，GPT-4o 从 96% 降至 63%，Gemini-2.5-Flash 从 96% 降至 89%），这种下降也十分明显；而较小的模型情况更为严重（例如，GPT-4o-mini 为 24%，Qwen3-8B 为 27%）。这些结果表明，无论模型容量如何，在长对话中维持全局约束仍是一项根本性挑战。

工具选择相比之下，商业模型表现出了很强的鲁棒性。GPT-4o 和 Gemini-2.5-Flash 等系统即使在多轮场景中，也保持了极高的准确率（≥97%）。然而，较小的开源模型则出现了明显的性能下降，尤其是在同一对话中涉及多个工具时。例如，Qwen3-32B 的准确率降至47%，Ministral-8B则降至37%。这表明，对于规模较大的模型而言，将输出明确 grounding 到固定工具集相对容易；但较小的模型难以在更高的对话复杂度下保持一致性。

实体抽取是各模型中鲁棒性最强的任务。在单轮预订中，几乎所有模型都取得了接近完美的准确率（96–100%）。即使在引入用户改变想法和干扰信息的多轮对话中，性能仍然相对较高（通常为84–89%，最低为79%）。这种韧性可能源于目标字段的结构化特性——（日期、时间、人数）通常以明确的数字或简短短语表达。因此，模型能够以有限的歧义可靠地捕捉最终槽位值。我们推测，如果抽取目标需要更丰富的上下文理解，例如将“那个有菠萝的披萨”这样的用户自由形式提及映射到规范菜单项（夏威夷披萨），性能下降可能会更加明显。

我们还注意到一些特定于模型的模式。**Mistral** 在指令遵循任务中的表现尤其弱，即使在单轮场景下也是如此（27-58%），这表明它难以遵守长度约束。另一方面，**Gemma**模型在工具选择方面展现出显著的鲁棒性，其多轮表现几乎与单轮持平。有趣的是，较大的模型（例如 GPT-4o、Gemma-3-27B、Qwen-32B）在单轮和多轮之间的表现差距小于较小的模型（例如 GPT-4o-mini、Mistral-8B），凸显了模型容量对于在长时程交互中保持性能的重要作用。

总体而言，这些发现进一步表明，尽管多轮对话会普遍降低准确率，但影响程度在很大程度上取决于任务类型和模型家族，而全局指令维持则是最具挑战性的维度。

详细错误分析

为了解可靠性下降的原因，我们按对话长度、任务复杂度和实体提取场景类型分析结果（完整表格见附录）。

对话长度本身并非导致指令遵循能力下降的主要因素。准确率虽有波动，但在较长对话中并未持续下降；在 10 轮对话时，准确率甚至达到 96% 的峰值。这表明，失败的原因不只是上下文长度，还可能来自特定干扰项或竞争性用户需求。

	GPT-4o	GPT-4o-mini	Gemini-2.5-Flash	Gemma-3-12B
单轮				
指令遵循	96	93	96	92
工具选择	100	100	100	100
实体抽取	100	96	100	92
多轮				
指令遵循	63	24	89	33
工具选择	99	93	97	98
实体抽取	86	84	89	79

	Qwen3-4B	Qwen3-8B	Qwen3-32B	Ministral-8B	Mistral-small-24B
单轮					
指令遵循	84	83	92	27	58
工具选择	100	100	100	99	100
实体抽取	97	98	100	100	100
多轮					
指令遵循	32	27	54	11	20
工具选择	58	89	47	37	71
实体抽取	83	88	89	88	89

表1：所有模型在单轮和多轮场景下各项任务的准确率（%）。

任务复杂度的影响更为明显：随着候选工具数量增加，工具选择的准确率急剧下降——从两个工具时的 98% 降至五个工具时的 64%。模型难以在多个干扰项中识别相关上下文，这表明真实的多智能体系统或工具使用系统面临着切实风险。

实体抽取错误因场景而异。日期 槽位始终表现最弱，反映出模型难以追踪时间信息。想法变化 类对话最容易出错（85%），而多次提及 类案例则相对稳健（91%）。在真实的预订任务中，时间变化或无关闲聊等对话干扰对可靠性的影响各不相同。

定性错误分析

代表性定性示例（完整结果见附录）展示了多轮场景中的典型失败模式。这些案例揭示了冗长且信息密集的提示、话题转换和误导性提及如何破坏对话一致性，并逐渐削弱任务可靠性。

在指令遵循案例中，经过几轮无关对话后，模型忽略了全局约束——尽管被要求将历史摘要控制在五句话以内，却生成了十三句话。这表明随着对话不断延长，模型遵循约束的能力会逐渐减弱，说明即使是风格或格式约束，在持续交互中也很脆弱。在工具选择任务中，话题混杂导致模型重复使用上一个工具（“股票”），即使用户接下来的查询明确需要另一个工具（“天气”）也是如此，这反映出模型过度依赖近期上下文，且意图重新评估能力不足。对于必须在不同领域间动态切换的现实助手而言，这种行为可能会产生累积影响。在实体抽取任务中，模型能够正确更新部分槽位，但会受到附近提及内容的干扰，从而覆盖最终预订时间。这说明短暂的上下文干扰会扰乱工作记忆，并削弱模型在对话轮次中跟踪结构化信息的可靠性。

总体而言，这些定性结果证实，性能下降并非仅由长度造成，而是源于特定的上下文冲突以及多轮对话中的记忆覆盖。

结论

我们研究了当任务被嵌入较长的、主题混杂的对话中时，对话可靠性如何下降。为了以可控且可复现的方式捕捉这一效应，我们引入了三项紧凑的多轮评估任务——全局指令遵循、工具选择和预订实体抽取——每一项都配有单轮对应任务。这一设计使我们能够分离可靠性下降的程度，并量化对话上下文如何影响大型商业大语言模型和较小的开源小语言模型的模型稳定性。我们发现，在简单场景中表现可靠的模型，当对话变得更长、更动态时，一致性往往会急剧下降，较小的模型尤其脆弱。

我们的贡献在于将多轮评估置于从业者部署对话系统时实际面临的场景中。与抽象基准不同，我们的任务强调实际可靠性因素——保持行为一致、处理多样化请求，以及随着时间推移维持结构化状态。该评估框架提供可复现的通过/失败指标和任务特定错误分析，从而为对话系统脆弱性出现的环节提供可操作的见解。这些发现凸显了将多轮可靠性视为模型评估核心维度的必要性，并指出应开发能够在真实对话条件下更好地保持鲁棒性的方法。

参考文献

Epstein, E. L.; Yao, K.; Li, J.; Bai, X.; and Palangi, H. 2024年。MMMT-IF：一项具有挑战性的多模态多轮指令遵循基准。arXiv:2409.18216。He, Y.; Jin, D.; Wang, C.; Bi, C.; Mandyam, K.; Zhang, H.; Zhu, C.; Li, N.; Xu, T.; Lv, H.; Bhosale, S.; Zhu, C.; Sankararaman, K. A.; Helenowski, E.; Kambadur, M.; Tayade, A.; Ma, H.; Fang, H.; and Wang, S. 2024年。Multi-IF：多轮和多语言指令遵循大语言模型基准测试。arXiv:2410.15553。Laban, P.; Hayashi, H.; Zhou, Y.; and Neville, J. 2025年。大语言模型在多轮对话中迷失。arXiv:2505.06120。Li, J.; Li, J.; Wang, Y.; Chang, Y.; and Wu, Y. 2025年。Struct-FlowBench：多轮指令遵循的结构化流程基准。arXiv:2502.14494。Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2023年。迷失在中间：语言模型如何利用长上下文。arXiv:2307.03172。Wang, X.; Wang, Z.; Liu, J.; Chen, Y.; Yuan, L.; Peng, H.; and Ji, H. 2024年。MINT：评估大语言模型在结合工具和语言反馈的多轮交互中的表现。arXiv:2309.10691。

详细定量结果

本附录提供了详细错误分析部分中引用的详细定量表格。结果按对话长度、工具数量和实体提取场景类型进行了细分。

表 2：不同对话轮数下的指令遵循准确率。

轮次	准确率
5	0.40
6	0.28
7	0.38
8	0.15
9	0.29
11	0.25

表 3：按对话长度划分的工具选择准确率。

轮次	准确率
6	0.76
7	0.74
8	0.82
9	0.74
10	0.96
11	0.78
12	0.80
13	0.75
14	0.81
15	0.75

表 4：按可用工具数量划分的工具选择准确率。

工具数量	准确率
2	0.98
3	0.82
4	0.74
5	0.64
6	0.71

表 5：按场景类型划分的实体抽取准确率。

场景类型	Date	Time	人物	全部匹配
想法变化	0.91	0.96	0.99	0.85
中间闲聊	0.91	0.97	0.97	0.86
多次提及	0.94	0.98	0.99	0.91

代表性定性示例

表 7 总结了定性错误分析部分提及的定性错误示例。这些案例展示了导致观测到的可靠性下降的主要失败模式，例如未能遵循指令、意图混淆以及受到无关提及的干扰。

表6：代表性定性错误案例。

上下文（摘录）	模型输出	标准答案
指令遵循 用户：从现在起，始终以至多5句话作答。 （在10轮无关闲聊之后） 用户：你能带我讲讲整个拜占庭帝国的兴衰全过程吗?	提供冗长而详细的说明——说明共 13 句话。	简洁的 4-5 句话总结。
工具选择 用户：AAPL 的股票价格是多少？截至——day? 用户：那巴黎现在的天气now?	[股票]	[天气]
实体抽取 用户：我想预订4月18日7点的三人桌pm. 用户：顺便说一下，我的朋友Sarah预订了昨天晚上6点。 用户：哦，其实改成4个人。	{”日期”: ”04/18”, ”时间”: ”下午7点”, ”人物”: ”4”}	{”日期”: ”04/18”, ”时间”: ”下午6点”, ”人物”: ”4”}

表7：代表性定性错误案例。模型经常（i）跟丢全局指令，（ii）在主题转换时选错工具，或（iii）被无关提及分散注意力。

统计显著性检验

为验证单轮与多轮场景之间观察到的性能差距并非由随机波动造成，我们针对每项任务进行了麦克尼马尔检验。检验采用了精确二项版本和连续性校正的 χ^2 近似，针对配对的单轮与多轮样本进行（每项任务 $n = 900$ ）。

Task	单轮准确率	多轮准确率	精确 p	χ^2 (cc)	p
指令遵循	0.764	0.396	1.7×10^{-61}	244.56	3.99×10^{-55}
工具选择	0.999	0.763	3.0×10^{-64}	210.00	1.37×10^{-47}
实体抽取	0.981	0.924	5.2×10^{-9}	31.65	1.85×10^{-8}

表8：每项任务基于900个配对样本的麦克尼马尔检验结果。单轮与多轮场景之间的所有差异均具有统计显著性。

用于数据生成的提示模板

本附录列出了用于生成三项评估任务合成对话的完整提示。我们分别提供了单轮和多轮场景的模板。为确保可复现性，所有实验均使用 GPT-5 并通过受控随机种子生成。此外，对话长度、主题构成和干扰密度均通过格式变量进行参数化，使我们能够在保持提示结构一致的同时，系统地控制实验条件。

指令遵循

单轮提示。

```
你是单轮指令遵循基准的数据生成器。
目标：
仅返回一个恰好包含一条用户消息的 JSON 对象（不包含助手消息）。
规则：
- 语言：英语。
- 这条用户消息必须以以下指令开头："始终用不超过 5 个句子作答。"
- 在同一条用户消息中，紧接该指令之后，围绕主题“{topic}”提出一个自然的问题。
- 该问题应自然地促使助手给出冗长、详细的回答（例如，要求完整概述故事、解释复杂主题或说明分步流程）。
- 不得包含任何助手消息。
- 仅输出具有以下确切结构的有效 JSON 对象：
{{
"messages": [{{"role": "user", "content": "<包含该指令和问题的单条用户消息>"}}]
}}
```

多轮提示。

```
你是一个多轮对话基准的数据生成器。
规则：
- 对话必须使用英语。
- 第一条用户消息必须是："From now on, always answer in at most 5 sentences"
- 对话总共应有 {turn} 轮，用户和助手交替发言。
- 在第一条指令之后，自然地围绕 {topic} 展开对话。
助手必须严格遵守指令（每次回复最多 5 句话）。
- 最后一轮必须由用户发言。
- 最后一条用户消息应自然地诱使助手给出冗长、详细的回答（例如："Can you tell me the whole story of Cinderella in detail?" 或 "Explain the entire process of how airplanes are built"）。
- 仅返回格式如下的有效 JSON 对象：
{{
"messages": [
{{"role": "user", "content": "From now on, always answer in at most 5 sentences ."}},
{{"role": "assistant", "content": "<助手消息 1，最多 5 句话>"}},
{{"role": "user", "content": "<用户消息 2>"}},
{{"role": "assistant", "content": "<助手消息 2，最多 5 句话>"}},
...
{{"role": "user", "content": "<诱使助手给出冗长回答的最后一条用户消息>"}}
]
}}
```

工具选择

单轮提示。

```
你是一个用于多轮工具选择基准测试的数据生成器。
可用工具：
```

4 {tools} 5 6 你的任务： 7 - 以 JSON 格式生成且仅生成一条用户消息（不包含助手回复）。 8 - 这条用户消息应是一项自然的请求，并且明确需要使用列表中的某一个工具。 9 - 这条消息的表达方式应自然且可以略带模糊，但仍须清晰到让人无需额外上下文即可选择正确的工具。 10 - "answer" 字段中只能填写该用户消息对应的正确工具名称。 11 - 输出必须严格遵循以下 JSON 格式： 12 13 {{ 14 "messages": [15 {{"role": "user", "content": "<单条用户请求>"}}16],17 "answer": "<其中一个工具>" 18 }} 19 20 不要包含任何解释，只返回 JSON 对象。

多轮提示1

你是一个用于多轮工具选择基准测试的数据生成器。 2 3 可用工具： 4 {tools} 5 6 你的任务： 7 - 以 JSON 格式生成恰好一个对话（包含 {turn} 个用户轮次）。 8 - 每一轮都只能来自用户（不得包含助手回复）。 9 - 对话应混合工具列表中的 {num_tools} 个主题，以增加工具选择的难度。 10 - 最后一轮必须明确需要使用其中一个工具。 11 - "answer" 字段只应填写最后一轮对应的正确工具。 12 - "mentioned_tools" 字段应列出对话中出现或与对话相关的所有工具（不只是最后一轮涉及的工具）。 13 - 用户话语应自然、口语化，并可通过混合主题有时起到误导作用。 14 - 输出必须严格遵循以下 JSON 格式： 15 16 {{ 17 "messages": [18 {{"role": "user", "content": "<用户消息 1>"}}, 19 ...20],21 "answer": "<其中一个工具>", 22 "mentioned_tools": ["<工具1>", "<工具2>", ...]23 }} 24 25 请勿添加任何解释，只返回 JSON 对象。

实体抽取

单轮提示。

1 你是一个用于单轮餐厅预订实体抽取基准的数据生成器。 2 3 你的任务： 4 - 以 JSON 格式恰好生成一条用户消息（不要助手回复）。 5 - 这条用户消息是关于预订餐厅餐桌的，并且必须包含： 6 - 明确提及**日期、时间和人数**（三者均为必需）。 7 - 日期必须以自然语言表达，且无需外部知识即可解析为固定日历日期。 8 可接受的示例包括明确日期（“3月15日”、“7月4日”、“10月21日”）或固定节日，例如“除夕”、“圣诞节”或“情人节”。 9 不允许使用相对时间指称，如“本周五”、“下周”，或可变日期节日，如“感恩节”。

10 - 时间必须具体且可解析（例如，“下午7点”、“正午”、“午夜”、“上午9点”）。 11 避免使用“下午”或“晚上”等模糊表述。 12 - 人数必须直接表达（例如，“一张4人桌”） 13 或通过无歧义的指称表达（例如，“我和我的父母” -> 3，“我们两人” ->2）。 14 不允许使用模糊的群体指称，如 “ 我的朋友们 ” 或 “ 几个人 ”。 15 - 加入无关闲聊（天气、食物偏好、旅行计划），以提高抽取难度。 16 - 信息可以在消息中以任意顺序出现。 17 18 仅返回符合此确切格式的有效 JSON 对象： 19 20 { 21 " messages": [22 {"role": "user", "content": "<单条用户消息>"}23],24 "answer": { 25 "date": "MM/DD", 26 "time": "H[am|pm]", 27 "people": "N" 28 } 29 } 30 31 - “answer”字段必须仅包含最终预订详情，并采用以下确切格式： 32 - 时间使用12小时制，带“am”/“pm”，小时为1-12（不加前导零）， 33 - 人数应为整数字符串（例如，“2”、“4”）。 34 - “answer”中的日期、时间和人数必须与消息中的最终预订意图一致。 35 36 不要包含任何解释，只返回该 JSON 对象。

多轮提示。

1 你是餐厅预订实体抽取基准的数据生成器。2 3 请严格按照 JSON 格式生成且仅生成一段对话。4 5 始终适用的全局约束（无论选择何种类型）： 6 - 对话包含 {turn} 轮，且每一轮都只能由用户发言（不得包含助手发言）。7 - 必须明确提及**日期、时间和人数**（三者缺一不可）。8 - 日期必须用自然语言表达，且无需借助外部知识即可转换为固定日历日期。9 可接受的示例包括明确日期（“3月15日”“7月4日”“10月21日”），或“除夕”“圣诞节”或“情人节”等固定节日。10 不允许使用“本周五”“下周”之类的相对时间指称，也不允许使用“感恩节”之类的可变日期节日。11 - 时间必须具体且明确（例如“晚上7点”“中午”“午夜”“早上9点”）。12 避免使用“下午”或“晚上”等模糊说法。13 - 人数必须直接表述（例如“订一张4人桌”），14 或通过明确无歧义的指称表达（例如“我父母和我” -> 3，“我们两个人” ->2）。15 不允许使用“我的朋友们”或“几个人”等模糊的群体指称。16 - 在各轮中变换日期、时间和人数的出现顺序（不能总是日期 -> 时间 -> 人数）。17 - 用户最后一轮不得用一个完整清晰的句子重述所有细节；18 最终预订必须通过整合分散在多轮中的信息推断得出。19 20 本示例的可选特性约束：21 -（想法变化）： 在对话中，日期、时间或人数至少变更 1~2 次。22 -（中间闲聊）： 在预订相关轮次之间穿插与预订无关的闲聊（天气、偏好、旅行等）。

23 -（多次提及）：偶尔提及为不同人物进行的多个预订，但答案中只能计入一个最终预订。