DynamicBench：评估大型语言模型中的实时报告生成

Jingyao Li[1], Hao Sun[2], Zile Qiao[2*], Yong Jiang[2],
Pengjun Xie[2], Fei Huang[2], Hong Xu[1], Jiaya Jia[3]

[1]The Chinese University of Hong Kong, [2]Alibaba Group,
[3]The Hong Kong University of Science and Technology

摘要

传统的大型语言模型（LLMs）基准测试通常依赖于通过讲故事或表达意见来进行静态评估，这无法捕捉到当代应用中实时信息处理的动态需求。为了解决这一局限性，我们提出了 DynamicBench，这是一个旨在评估 LLMs 存储和处理最新数据能力的基准测试。DynamicBench 利用双路径检索管道，结合网络搜索与本地报告数据库。它需要特定领域的知识，确保在专业领域内准确生成报告。通过在提供或不提供外部文档的情景中评估模型，DynamicBench 有效衡量了它们独立处理最新信息或利用上下文增强的能力。此外，我们引入了一个先进的报告生成系统，能够管理动态信息综合。我们的实验结果证实了我们方法的有效性，该方法在无文档和有文档辅助的场景中分别比 GPT4o 高出 7.0%和 5.8%，达到了最先进的性能水平。代码和数据将公开发布。

1 介绍

近年来，大型语言模型（LLMs）彻底改变了自然语言处理领域，在从语言生成到各种领域的语境理解等任务中表现出卓越的技能。然而，传统的基准测试仍然局限于静态评估，通常依赖于讲故事或表达意见。这种静态、主观的评估标准无法捕捉实时信息处理的动态特性，这对于理解大型语言模型 (Wu et al., 2025; Que et al., 2024) 的真实能力至关重要。

针对这些限制，我们引入了 DynamicBench，一个旨在评估 LLMs 获取和处理实时数据能力的基准测试。以其通过网络搜索和数据库查询检索到的当代信息为特色，DynamicBench 要求模型具备最新的知识以提供准确的回答。利用双路径检索管道，DynamicBench 结合了本地报告数据库与网络搜索，确保能够访问全面的数据来进行彻底的报告评估。DynamicBench 评估了广泛领域的最新动态，涵盖了诸如科技与科学、经济与环境、文化和健康和国际与政治等关键类别。通过提供或不提供外部文档的两种场景，DynamicBench 评估了模型存储知识或有效处理最近外部信息的能力。在专业领域内精确数据收集的要求保证了评价过程的准确性和客观性，弥补了当前方法论在客观和实时评估方面的不足。

除了基准测试本身，我们的贡献还包括一个强大的报告生成解决方案，能够应对动态信息生成带来的复杂挑战。我们的系统从基于查询的报告规划开始，并使用本地和在线数据的双路径检索管道进行查询生成和资源聚合。该系统自我评估是否需要进一步收集信息，并确保充分的信息采集，为整合表格和图表以增强清晰度的详细报告撰写提供支持。最终，它输出一个全面、连贯的报告，反映最新数据。实验结果证明了我们方法的有效性。我们在两种条件下评估 LLMs：无文档辅助和有文档辅助，并

*Corresponding Author

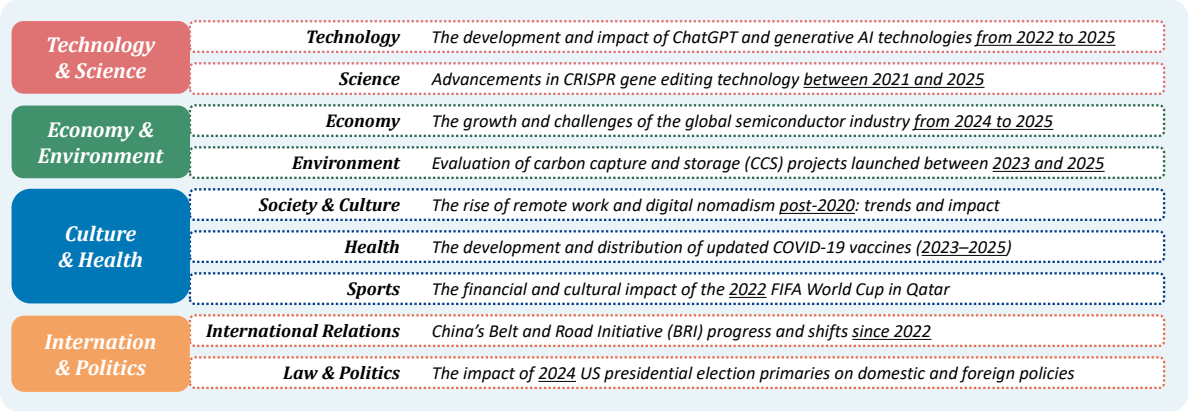


图 1: 查询示例涵盖四个主要类别：科技与科学、经济与环境、文化与健康和国际与政治，每个类别都有多个子类别。

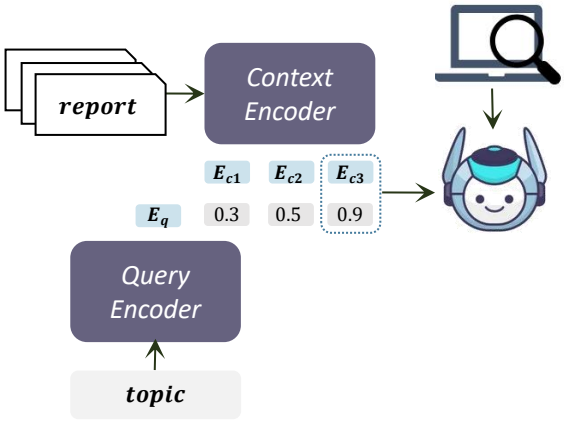


图 2: 双路径检索报告生成系统，该系统结合了从本地财务报告数据库增强的检索生成（RAG）和网络搜索来收集信息。相关信息被输入到大型语言模型中以进行综合报告生成。

分析这两种情况下它们在不同领域的表现。我们的方法在多个指标上展示了最先进的性能，分别比 GPT4o 高出 7.0%和 5.8%。

总之，我们的贡献如下：

1. 我们引入了 DynamicBench ，这是一个新颖的基准测试，根据实时信息获取和处理能力评估大语言模型，利用结合本地和在线数据源的双路径检索系统。
2. 我们开发了一个全面的报告生成系统，该系统规划、搜索并撰写详细报告，确保整合最新的信息以实现准确和连贯的文档记录。
3. 我们通过实验结果展示了我们方法的先进能力，与领先的 LLMs 相比，我们的方法达到了最先进的性能，在多个指标上都有显著提升。

2 相关工作

2.1 编写基准测试

在评估大型语言模型（LLMs）方面的最新进展导致了多个基准测试的创建，旨在评估不同方面的语言生成和理解能力。LongBench-Write (Bai et al., 2024) 专注于了解模型在处理复杂写作任务中的能力。HelloBench (Que et al., 2024) 扩展了评估工作，通过将长文本生成分类为开放性问答和启发式文本生成等不同任务。EQ-Bench (Paech, 2024) 引入了一种情感智能的评估方法，通过评估 LLMs 理解和预测对话中情绪强度的能力来进行。WritingBench (Wu et al., 2025) 提供了一个全面的跨领域和子领域的评估，包括创意写作和技术写作。这些传统方法主要关注讲故事或表达意见，采用静态和主观的评估标准。相比之下，我们的系统不仅提供了一个涵盖广泛主题并评估写作各个方面的一体化框架，而且还利用实时网络搜索和数据库查询来访问最新信息。因此，我们的系统评估了模型处理和有效使用实时信息的能力。此外，我们的基准测试需要在专业领域内构建精确报告，从而确保所使用的数据的准确性和

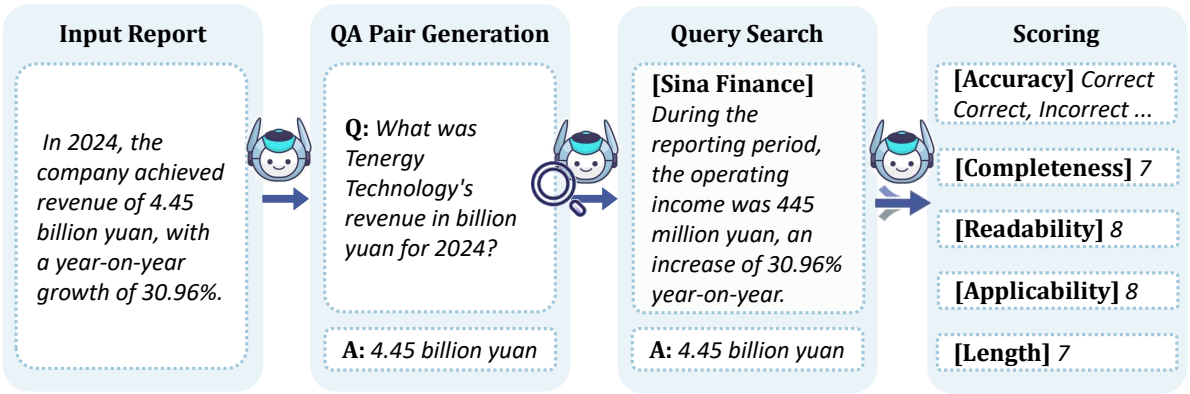


图 3: 评估系统流程始于从输入报告中提取的关键细节生成问题和答案（Q&A）对，这些对作为查询用于双路径检索策略。该策略涉及同时在线和本地财务报告数据库中搜索以收集相关信息。系统通过将报告数据与检索到的信息进行比对来评估每个 Q&A 对的准确性，并计算准确性。报告的完备性、可读性、适用性和长度也基于检索到的信息进行评估。

客观性。这些特点使我们的基准能够在当前基准中关于客观和实时评估方面存在的差距。

2.2 长上下文能力的大型语言模型

大型语言模型（LLMs）如 Claude-3 (Anthropic, 2023), DeepSeek-R1 (DeepSeek-AI, 2025), DeepSeek-v3 (DeepSeek-AI et al., 2025), GPT-4o (OpenAI et al., 2024), 和 Qwen-2.5 (Qwen et al., 2025) 在包括理解和生成复杂语言任务在内的多个领域展示了卓越的能力。这些模型作为众多应用的基础工具，但在生成扩展输出或遵循复杂的任务约束方面常常面临限制。LongWriter (Bai et al., 2024) 通过提出 AgentWrite，一个基于代理的管道，解决了当前 LLMs 的输出长度限制问题，使模型能够生成超过 20,000 字的一致性输出。Suri (Pham et al., 2024) 引入了一种多约束指令遵循方法，用于生成长篇文本。它可以生成质量持续稳定且符合约束条件的显著更长的文本。相比之下，我们的工作通过有效生成具有增强连贯性和质量的扩展内容超越了之前的努力。

3 方法论

为了解决动态信息生成带来的挑战以及对准确报告构建的需求，我们的方法论围绕着基准的发展和一个稳健的系统解决方案展开。在 section 3.1 中，我们介绍了一个旨在评估大语言模型获取和处理实时数据能力的基准。在 section 3.2 中，我们提出了我们的报告生成系统解决方案。

3.1 动态基准测试

传统上，基准测试 (Wu et al., 2025; Que et al., 2024) 依赖于讲故事或表达意见，由于其静态性质，这些方法不具有时间敏感性。相比之下，我们的基准，如 Fig. 1 所示，需要通过网络搜索和数据库查询检索当代的、有时间敏感性的信息。这种方法要求拥有最新的特定领域知识以提供准确的回答，从而评估当前模型获取和处理实时信息的能力。此外，与传统的主观评价不同，我们的基准要求收集数据来构建专业领域的报告，确保评估的准确性和客观性。这些特性使我们的基准能够缩小目前在客观和实时评估方面的差距。我们的基准包括以下类别：

- **科技与科学**：技术和科学。
- **经济与环境**：经济和环境。
- **文化和健康**：社会与文化，健康和体育活动。
- **国际与政治**：国际关系 和 法律与政治。

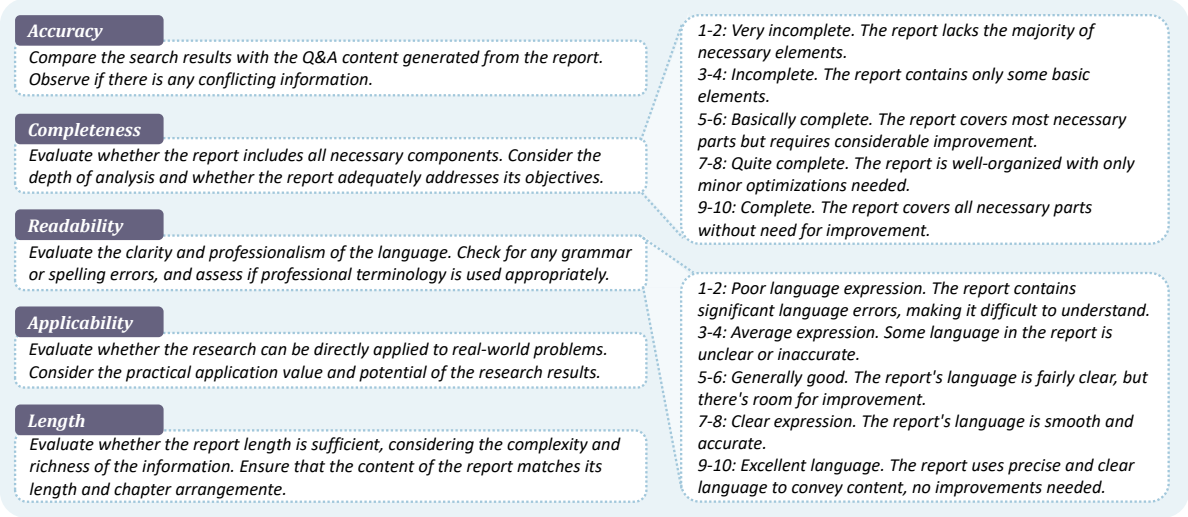


图 4: 报告分析的评估标准，重点关注五个不同的指标：准确性, 完备性, 可读性, 适用性和长度。每个标准都附有详细描述，以指导评估者比较搜索结果，检查分析深度，评估语言清晰度和专业性，评价实际应用性，并验证报告长度相对于内容丰富性的充分性。右侧的面板举例说明了从 1 到 10 的评分尺度对于完整性和可读性的具体指南，每个分数如何反映报告的质量和一致性。

3.1.1 信息检索过程

为了构建我们的本地报告数据库，我们从 AnnualReport[1] 获取了来自 10,338 家全球公司的 148,589 份年度报告。这些报告涵盖了经济、环境、技术、科学、文化、健康、法律、政治等各种领域。我们利用检索增强生成（RAG）方法进行信息检索，如图 Fig. 2 所示。

该过程涉及使用上下文编码器对本地报告数据库进行编码，以及使用查询编码器对接收的查询进行编码。每个报告和查询都会被转换成嵌入表示，分别用 $\boldsymbol{E}_c$ 表示上下文，用 $\boldsymbol{E}_q$ 表示查询。这些嵌入之间的相似性通过余弦相似度来计算，定义为：

$$\text{Similarity}(\boldsymbol{E}_c, \boldsymbol{E}_q) = \frac{\boldsymbol{E}_c \cdot \boldsymbol{E}_q}{\|\boldsymbol{E}_c\| \|\boldsymbol{E}_q\|} \quad (1)$$

系统有效地提取相似度分数最高的报告块作为最相关信息。此过程在数学上表示为选择块 $\boldsymbol{B}^*$，使得：

$$\boldsymbol{B}^* = \text{argmax}_i \ \text{Similarity}(\boldsymbol{E}_{c_i}, \boldsymbol{E}_q) \quad (2)$$

[1] https://www.annualreports.com

我们利用双重路径检索管道，通过本地报告数据库和网络搜索获取信息。这种全面的方法确保我们的系统能够利用可用数据生成稳健且全面的报告。

3.1.2 评估过程

如图 Fig. 3 所示。我们评估过程的初始阶段包括从输入报告中提取关键信息以生成问题和答案（Q&A）对。这些对构成了后续信息检索的基础，在此过程中，将使用一种双路径检索策略来应用由这些对衍生出的查询。该策略利用网络搜索和本地财务报告数据库收集全面的信息。一旦检索完成，我们的系统会评估多个指标，如图 Fig. 4 所示。这些指标包括：

准确性。 每个问答对的准确性通过评估报告数据与检索到的信息之间的一致性来确定；如果搜索数据证实了问答内容或未发现差异，系统将其标记为正确。反之，如果相关的内容缺失或检测到冲突，可能会被标记为无法确定或错误。通过计算所有查询的平均准确性来确定报告的最终准确性指标。

完备性。 该指标评估报告是否包含所有必要元素并充分实现其目标。评估者使用评分标准

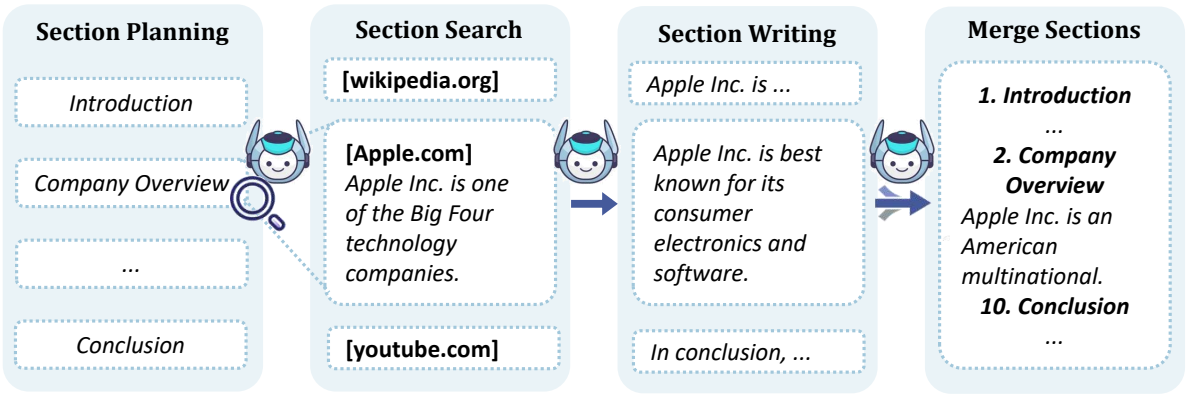


图 5: 主题苹果公司 *2021* 年财务业绩回顾的报告组成工作流程：该四步过程始于部分规划，在此阶段根据研究主题确定主要章节标题。接下来是章节搜索阶段，涉及对每个部分进行精确查询，从本地和在线来源收集相关证据。章节写作利用收集到的证据创建详细的分析文本，并附有表格和图表。最后，合并部分将所有章节分析整合成一份连贯的报告，优化叙述流程并输出完整文档。

来确定完整度，从非常不完整（1-2 分）到完全完整（9-10 分）。

可读性。 本标准评估报告语言的清晰度和专业性，检查语法和拼写错误以及专业术语的恰当使用。可读性从语言表达差（1-2 分）到语言运用优秀（9-10 分）。

适用性。 该指标衡量研究成果的实际应用价值，评估报告是否可以直接有助于解决现实世界的问题。适用性从较差的适用性（1-2 分）到显著的应用价值（9-10 分）。

长度。 该标准评估报告的长度是否充分涵盖了所呈现信息的复杂性和丰富性。长度评分从极不充分（1-2 分）到非常充分（9-10 分），考虑到每个章节内容的适宜性。

3.2 报告生成过程

在处理复杂的报告生成挑战时，我们的系统提供了一种结构化的方法论以实现高质量的输出，如 Fig. 5 所示。

部分规划。 此初始阶段涉及根据研究主题建立主要部分标题。例如，在回顾苹果公司 2021 年的财务表现时，确定了如引言、公司概览和结论等部分以逻辑性地组织报告。

章节搜索。 每个部分，模型最初生成 K 查询以检索相关数据和见解。这些查询用于搜索本地数据库和在线资源，并汇总检索到的内容。然后，模型对收集的信息进行自我评估，以确定是否足以起草该部分内容。如果内容被认为不足，则会生成并执行额外的查询以填补信息中的任何空白。这个迭代过程将持续进行，直到获取足够的数据，使系统能够进入下一阶段。

章节写作。 利用收集到的证据，系统为每个部分生成详细的分析文本。此阶段包括表格和图表的整合，提升报告的信息质量和视觉清晰度。

合并部分。 最后一步涉及将所有开发的部分整合成一份连贯的报告。系统优化叙述流程，确保文档对研究主题进行了全面、连贯的分析，并以最终准备发布的输出结论。

这一系统方法确保了全面、基于数据的报告制作，以生成高质量文档。

4 实验

基准模型。 基线大语言模型包括 Claude3.7 (Anthropic, 2023)、DeepSeek-R1 (DeepSeek-AI, 2025)、DeepSeek-v3 (DeepSeek-AI et al., 2025)、GPT4o (OpenAI et al., 2024) 和 Qwen-72B (Qwen et al., 2025)。此外，我们引入了能力增强的模型，如 Long-

Models	Acc.	Comp.	Read.	App.	Len.	Average
无文档						
DeepSeek-R1 (DeepSeek-AI, 2025)	40.8	62.0	77.6	69.3	52.3	60.4
DeepSeek-v3 (DeepSeek-AI et al., 2025)	44.1	65.9	77.4	69.9	64.6	64.4
Qwen2.5-72B-Instruct (Qwen et al., 2025)	49.3	61.3	75.8	68.1	60.8	63.1
GPT4o (OpenAI et al., 2024)	58.2	65.7	77.3	70.4	66.0	67.5
Claude3.7-Sonnet (Anthropic, 2023)	55.0	64.7	**79.0**	71.3	64.5	66.9
Suri (Pham et al., 2024)	43.9	45.5	62.6	63.0	43.0	51.6
LongWriter (Bai et al., 2024)	68.0	45.4	62.4	62.8	41.5	56.0
与文档						
DeepSeek-R1 (DeepSeek-AI, 2025)	15.3	47.0	53.0	64.4	42.9	44.5
DeepSeek-v3 (DeepSeek-AI et al., 2025)	59.9	69.4	77.2	70.1	68.3	69.0
Qwen2.5-72B-Instruct (Qwen et al., 2025)	63.6	60.3	75.4	66.8	60.4	65.3
GPT4o (OpenAI et al., 2024)	63.4	65.6	77.3	70.5	67.0	68.7
Claude3.7-Sonnet (Anthropic, 2023)	69.3	65.9	**78.7**	71.2	66.3	70.3
Suri (Pham et al., 2024)	51.7	40.2	57.2	60.9	37.2	49.5
LongWriter (Bai et al., 2024)	45.0	30.1	40.2	47.1	26.8	37.8
Ours	**74.8**	**73.7**	78.0	**71.7**	**74.4**	**74.5**

表 1: 评估大语言模型的指标包括多个方面，如准确性（Acc）、完整性（Comp）、可读性（Read）、适用性（App）和长度（Len）。带有 Doc 和 w/o Doc 表示是否向模型提供了相关文档。最佳和次佳结果分别用加粗和下划线格式突出显示。

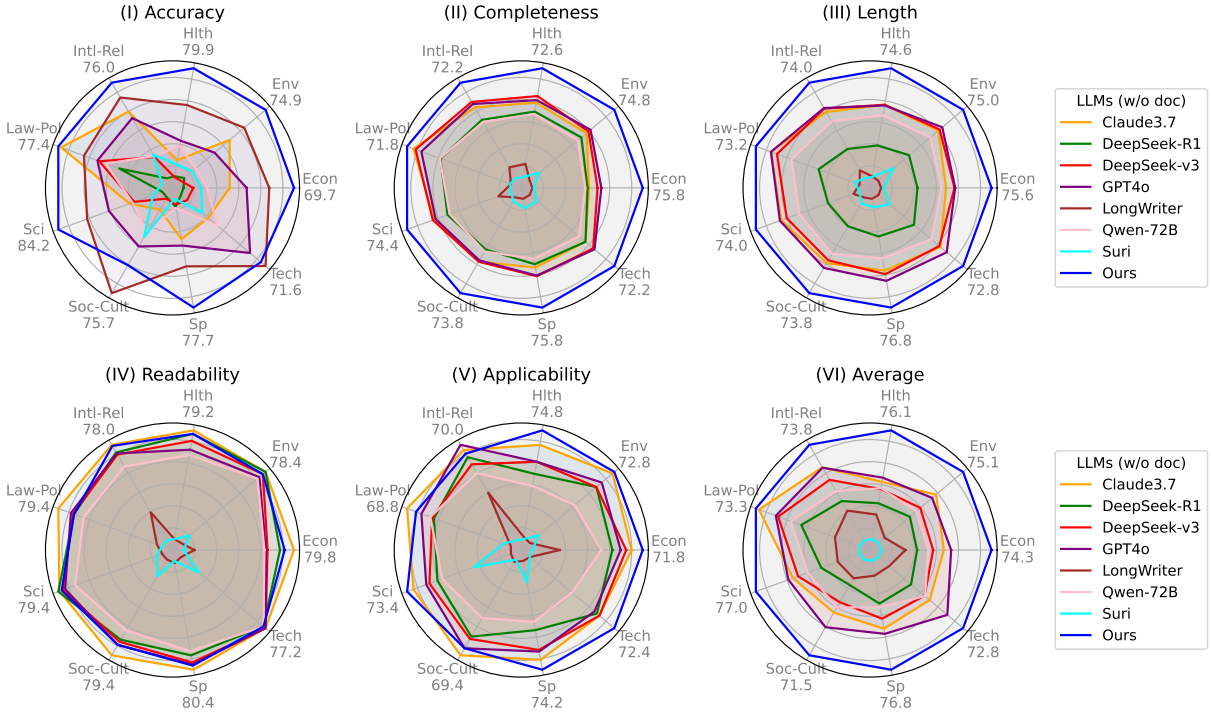
Writer (Bai et al., 2024) 和 Suri (Pham et al., 2024)。LongWriter 利用独特的方法论，例如 AgentWrite 管道，以促进超过 20,000 字连贯文本的生成。类似地，Suri 采用多约束指令遵循策略来生成显著更长的文本，同时确保质量和符合约束。

评估。 为了评估当前语言模型处理动态和实时数据的能力，我们采用了新开发的基准测试 DynamicBench 。该基准测试通过专注于获取和分析时效信息超越了传统方法。利用网络搜索和数据库查询，我们挑战模型展示它们在处理最新领域特定查询方面的熟练程度。这种方法全面评估了模型动态整合最新信息并跨各种专业领域构建准确报告的能力。评估维度包括准确性，完备性，可读性，适用性和长度。

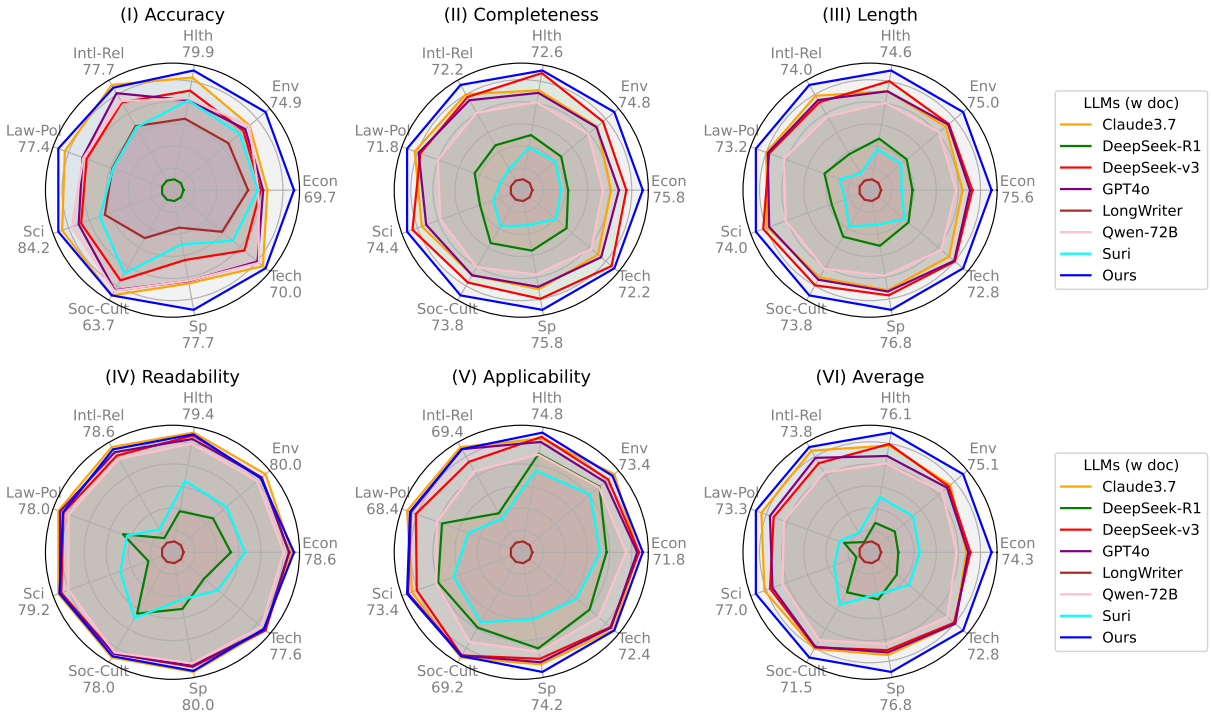
4.1 结果

在 Tab. 1 中，我们展示了我们的方法与基线模型在两种条件下评估的结果：**无文档**和**带有文档**。**无文档**设置涉及基线大语言模型在没有外部文档辅助的情况下进行回应，而**使用文档**设置允许它们利用我们系统的双路径检索结果。这些设置是为了评估每个模型独立处理信息的能力与利用额外上下文能力的对比。

无文档的 LLMs 我们的方法在所有维度上展示了新的最先进性能。在准确性方面，我们的模型达到了 74.8%，比 GPT4o 高出 16.6%。该类别的当前 SOTA 是能力增强模型 LongWriter，其达到了 69.3%。对于完备性，我们的方法获得了 73.7%的分数，超过了 DeepSeek-v3 8.1%。在评估可读性时，Claude3.7-Sonnet 以 78.7%



(a) 不带文档的 LLMs



(b) 具备文档的 LLMs

图 6: 各种系统和大语言模型的性能包括 Claude3.7 (Anthropic, 2023), DeepSeek-R1 (Anthropic, 2023), DeepSeek-v3 (DeepSeek-AI et al., 2025), GPT4o (OpenAI et al., 2024), LongWriter (Bai et al., 2024), Qwen-72B (Qwen et al., 2025), Suri (Pham et al., 2024)。评估指标包括准确性、完整性、长度、可读性、适用性和平均性能。每个模型都通过一系列主题进行评估，例如经济（Econ）、环境（Env）、健康（Hlth）、国际关系（Intl-Rel）、法律和政治（Law-Pol）、科学（Sci）、社会和文化（Soc-Cult）、体育（Sp）和技术（Tech）。

的成绩领先，紧随其后的是我们的模型，得分为 78.0%。在适用性方面，我们的模型展示了71.7%的得分，略胜于 Claude3.7-Sonnet。关于长度，我们的模型以 74.4%的成绩超越了竞争对手，比下一个最好的模型 GPT4o 高出 7.4%。在五个维度中，我们的方法平均得分为 74.5%，超过了当前 SOTA GPT4o 的平均分 7.0%。

带文档的 LLMs 具备相关文档的访问权限，我们的方法在所有评估指标中继续展示出最先进的性能。在准确性方面，我们的模型取得了 74.8%的出色成绩，超过了当前 SOTA Claude3.7-Sonnet 5.5%。对于完备性，我们的方法得分 73.7%，超过了 DeepSeek-v3 的成绩 4.3%。Claude3.7-Sonnet 在可读性的表现领先，得分为 78.7%，而我们的模型紧随其后，得分为 78.0%。我们的模型展示了卓越的适用性，得分 71.7%，略微超过了 Claude3.7-Sonnet。在长度方面，我们的模型表现出色，得分为 74.4%，明显优于 DeepSeek-v3 的成绩，领先了 6.1%。总体而言，在五个维度中，我们的方法取得了平均分 74.5%，显著高于 Claude3.7-Sonnet 和 GPT4o 的 4.2%和 5.8%。

比较。 带有和不带有文档访问权限的模型评估显示出显著的性能差异。对于如 Qwen2.5-72B-Instruct、DeepSeek-v3、GPT4o 和 Claude3.7-Sonnet 这样的通用大语言模型，当提供相关文档时，其性能普遍有了极大的提升。这种增强突显了大语言模型有效利用外部上下文的能力。相反地，对于像 Suri 和 LongWriter 这样经过能力增强的模型，在包含文档输入的情况下观察到性能下降。这表明这些优化用于生成扩展文本的模型可能在提供额外文档时牺牲了一部分理解长上下文的能力。这种倾向可能导致引入外部数据后可读性和完整性降低。此外，无论是通过人类反馈（RLHF）进行微调的 DeepSeek-v3 和 Suri 都表现出这一模式，暗示它们的训练方法可能更侧重于生成方面而非上下文理解。

4.2 类别级分析

在 Fig. 6 中，我们展示了各个领域 LLMs 和系统的详细结果。对于**没有文档的大型语言模型**，尽管以前方法的平均结果与我们的结果相差显著，在某些领域仍然可以取得更好的效果。例如，LongWriter 在社会与文化和技术领域的准确性略高于我们，而 Claude3.7 在法律与政治和社会与文化领域的应用性略好。这可能是因为这些模型在训练过程中形成了偏好，可能是由于包含了网络搜索或现有数据库中未包含的具体知识。很明显，**具有文档的 LLMs** 的平均结果与**没有文档的 LLMs** 相比有显著提高，这源于补充的外部信息增强了大语言模型的内在知识。虽然**带有文档的大型语言模型**的结果已经缩小了与我们系统的差距，但总体上并未超越我们的系统，这可以归因于**具有文档的 LLMs** 和我们的系统都使用了相同的双路径检索信息。然而，我们的方法通过系统的方法有效利用了这些信息。

5 结论

本工作通过引入 DynamicBench ，一种旨在评估实时信息获取和处理能力的动态基准测试，超越了传统用于评估大规模语言模型（LLMs）的标准。利用结合本地报告数据库与网络搜索的双路径检索系统，DynamicBench 提供了跨不同领域的全面且客观的评估。此基准要求模型展示特定领域的知识，确保生成准确的报告。此外，我们开发了一个先进的报告生成系统，能够处理动态信息综合中的复杂性。通过系统的规划、查询生成和资源整合，该系统集成最新的信息以生成详细连贯的报告，反映最新数据趋势。我们的实验结果强调了其有效性，展示了在各种场景中超越现有模型如 GPT4o 的最先进性能。

限制

虽然 DynamicBench 通过集成实时数据检索和处理在评估大语言模型方面取得了进展，但仍存在一些限制。首先，依赖于网络搜索和本地报告数据库意味着基准的有效性取决于这些外部来源的质量和可访问性。可用数据中的差异或偏差可能影响评估结果的准确性和客观性。此外，考虑的文档范围可能无法捕捉到特定领域所需的完整背景知识。该基准可能无法全面评估需要高度专业见解和专业知识的专业领域的理解深度。这些限制突出了潜在改进的领域，为未来的工作铺平了道路，专注于增强数据整合策略，并扩展领域覆盖范围以进一步推进大语言模型评估和报告生成方法。

更广泛的影响

随着 AI 模型越来越能够处理实时信息，关于这些技术的伦理使用和潜在滥用的问题也随之而来。快速生成详细、连贯报告及实时数据整合的能力增加了部署 LLMs 用于制造误导或有偏见内容的风险。研究人员和开发者必须优先考虑减轻此类风险。通过在 AI 实践中培养透明度和问责制，我们可以确保我们的工作产生积极影响，同时遏制可能的危害或滥用。最终，我们的努力旨在为利益相关者提供增强的工具，以负责任地应对现代信息环境的复杂性。

人工智能辅助披露

作者们整合了大语言模型以辅助起草本手稿的部分内容。在文本初次创建后，作者们彻底审查并精炼了材料，确保其准确性和完整性，并且他们对已发表的工作承担全部责任。

References

Anthropic. 2023. Claude 3 model card. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. Accessed: 2023-10-05.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longwriter: Unleashing 10,000+ word generation from long context llms. *Preprint*, arXiv:2408.07055.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, ukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, ukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Samuel J. Paech. 2024. Eq-bench: An emotional intelligence benchmark for large language models. *Preprint*, arXiv:2312.06281.

Chau Minh Pham, Simeng Sun, and Mohit Iyyer. 2024. Suri: Multi-constraint instruction following for long-form text generation. *Preprint*, arXiv:2406.19371.

Haoran Que, Feiyu Duan, Liqun He, Yutao Mou, Wangchunshu Zhou, Jiaheng Liu, Wenge Rong, Zekun Moore Wang, Jian Yang, Ge Zhang, Junran Peng, Zhaoxiang Zhang, Songyang Zhang, and Kai Chen. 2024. Hellobench: Evaluating long text generation capabilities of large language models. *Preprint*, arXiv:2409.16191.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang,

Mengyue Wu, Qin Jin, and Fei Huang. 2025. Writingbench: A comprehensive benchmark for generative writing. *Preprint*, arXiv:2503.05244.