

使用自控记忆系统释放大规模语言模型的无限长度输入能力

Xinnian Liang^{1*}, Bing Wang^{1*}, Hui Huang^{3†}, Shuangzhi Wu^{2‡}

Peihao Wu², Lu Lu², Zejun Ma² and Zhoujun Li^{1‡}

¹State Key Lab of Software Development Environment, Beihang University, Beijing, China

²ByteDance AI Lab, Beijing, China

³Harbin Institute of Technology, Harbin, China

{ xnliang,bingwang,lizj } @buaa.edu.cn; { wufurui,mazejun } @bytedance.com;

Abstract

大型语言模型 (LLM) 受限于无法处理冗长的输入。为了解决这一局限性, 我们提出了自控记忆 (SCM) 系统来为大规模语言模型释放无限长度的输入能力。我们的 SCM 系统由三个关键模块组成: 语言模型代理、内存流和内存控制器。语言模型代理迭代处理超长输入并将所有历史信息存储在内存流中。记忆控制器为代理提供长期记忆 (存档记忆) 和短期记忆 (闪存) 以生成精确和连贯的响应。控制器确定应激活存档内存中的哪些内存以及如何将它们合并到模型输入中。我们的 SCM 系统可以与任何 LLM 集成, 使他们无需任何修改或微调即可处理超长文本。实验结果表明, 我们的 SCM 系统使未针对多轮对话优化的 LLM 能够实现与 ChatGPT 相当的多轮对话能力, 并且在涉及超长文档摘要或长期对话的场景中优于 ChatGPT。此外, 我们将提供一个测试集, 涵盖常见的长文本输入场景, 用于评估 LLM 处理长文档的能力。^{1 2}

1 介绍

最近, 大规模语言模型 (LLM) 因其在各种任务中的卓越表现而备受关注 (Brown et al., 2020a; Zeng et al., 2023; Ouyang et al., 2022; Thoppilan et al., 2022; Workshop et al., 2023; Muennighoff et al., 2022)。LLM 强大的基础能力是通过对大量文本语料库进行大规模预训练 (例如, In-Context Learning (Brown et al., 2020a)、Chain-of-Thoughts (Wei et al., 2022c,b) 等) 实现的, 这是它们成功的一个促成因素。指令调整 (Raffel et al., 2020; Wei et al., 2022a; Chung et al., 2022) 帮助 LLM 理解自然语言任务描述, 而人类反馈强化学习 (RLHF) (Schulman et al., 2017; Stiennon et al., 2020; Bai et al., 2022) 使生成的文本与人类偏好保持一致。LLMs 的综合能力有效地打破了自然语言处理任务之间的界限, 为 LLMs 的应用和研究方向带来了无限可能。

* The authors contribute equally and the order is random.


† Contribution during internship at ByteDance Inc.

‡ Corresponding Author

¹ 工作中。

² <https://github.com/wbbeyourself/SCM4LLMs>

[User]: What are your hobbies?
[System]: As an AI language model, I don't have hobbies like humans do. However, I am programmed to assist with various tasks and answer questions. Is there anything specific you need help with?
[User]: Tell me my hobby: swimming, climbing mountain and traveling. (50+ turns of dialogue about hobby and other topics is omitted here.)
[User]: We talk about hobbies again. What do you think is my hobby?

[ChatGPT]:  As an AI language model, I don't have access to your personal information or interests, so I can't say for sure what your hobby is. However, I can suggest some common hobbies that people enjoy such as reading, writing, painting, playing sports, cooking, gardening, playing musical instruments, and many more. You can explore different hobbies and find the one that brings you joy and fulfillment.


[Our SCM System]:  Based on our previous conversation, your hobbies include swimming, climbing mountains, and traveling. Is there anything related to those hobbies that you need assistance with?

Figure 1: 比较 ChatGPT 和我们的 SCM 系统的示例。橙色文本表示用户当前的问题, 粗体文本表示相关历史信息, 红色文本表示 ChatGPT 的错误响应, 绿色文本表示我们系统的正确响应。

大型语言模型 (LLM) 具有众多优点, 但它们的实用性受到两个主要因素的阻碍: 最大输入长度和预训练阶段自我注意的计算复杂性 (Wang et al., 2020)。尽管某些模型 (Press et al., 2022; OpenAI, 2022) 能够处理长输入, 但它们可能仍难以在异常冗长的文本中捕获关键的上下文信息。如图 1 所示, 即使是 ChatGPT³ 也会因为历史噪声的积累而错过前面文本的重要上下文, 历史噪声指的是可能阻碍理解的不相关或过时的信息。

为了解决这一限制, 我们提出了自控记忆 (SCM) 系统, 它使大型语言模型 (LLM) 无需任何修改或额外训练即可处理无限长度的文本。输入被分成多个部分并作为观察 (输入) 馈送到 LLM。SCM 对 LLM 进行了扩展, 增加了长期记忆 (归档记忆)、短期记忆 (闪存) 和记忆控制器。存档内存保留所有历史信息, 而闪存捕获前几轮的实时内存信息。内存控制器决定何时以及如何引入归档信息, 使 LLM 能够在不牺牲任何基本信息的情况下高效处理超长文本。

³ 在本研究中, 我们使用 OpenAI gpt-3.5-turbo-0301。

为了评估我们系统的性能,我们将 SCM 与非对话优化的 LLM 集成,并成功模拟 ChatGPT。我们的研究表明,我们的系统在处理超长输入或对话方面优于 ChatGPT。对于摘要任务,我们生成整个归档内存的分层摘要,直到摘要长度满足用户的规范。通过将来自前面文本的信息合并到内存中的本地摘要中,我们的方法保留了原始内容之间的相关性,这与直接生成整个文本的分层摘要的传统方法相反。此外,我们的工作仍在进行中,我们计划发布一个专为长文本任务设计的综合评估数据集,以及标准化的人工评估,以评估不同方法的有效性。

2 相关工作

大规模语言模型。 Large-scale Language Models (LLMs) 是在大量文本数据上训练的语言模型,使用 Transformer (Vaswani et al., 2017) 架构作为它们的基础。最早的基于 Transformer 的预训练语言模型是 GPT-1 (Radford et al., 2018)。随后, GPT-2 (Radford et al., 2019) 和 GPT-3 (Brown et al., 2020b) 逐渐增加参数大小。GPT-3 规模最大,有 175B 个参数,还有 emergent abilities (Wei et al., 2022b,c), 引起了学术界和工业界的关注。

从那以后,出现了很多 LLM, 包括 LAMBDA (Thoppilan et al., 2022), PaLM (Chowdhery et al., 2022), OPT (Zhang et al., 2022a), LLaMA (Touvron et al., 2023), BLOOM (Workshop et al., 2023), Galactica (Taylor et al., 2022), and Pangu (Zeng et al., 2021; Ren et al., 2023)。ChatGPT (OpenAI, 2022) 是这一系列研究中最引人注目的作品之一,引起了业界的广泛关注,被认为是通用人工智能 (AGI) 的里程碑,它基于 InstructGPT (Ouyang et al., 2022) 并针对多轮对话能力进行了优化。ChatGPT 取得了卓越的性能,超越了 NLP 任务之间的界限。然而,当前的 LLM, 包括 ChatGPT, 在处理涉及极长输入的任务时面临着很大的限制。

长文本序列处理。 处理长文本序列一直是自然语言处理任务中的一个持续挑战。随着预训练和 LLM 的出现,这个问题变得更加突出,因为预训练期间的固定输入长度和在预训练阶段扩展它的高成本限制了处理更长输入的能力。现有的解决方案主要是在预训练时更换 Attention 结构以降低计算成本和扩大预训练序列长度 (Beltagy et al., 2020; Zaheer et al., 2021; Guo et al., 2022; Phang et al., 2022; Dong et al., 2023)。另一种替代方法 (Press et al., 2022) 在预训练期间使用特殊的位置编码,使模型能够

学习相对位置并在推理期间处理更长的输入文本。

然而,这些方法的普遍性和它们对下游任务的影响仍然不确定。在长文本摘要领域,有很多行之有效的办法。Wu et al. (2021); Zhang et al. (2022b); Cao and Wang (2022) 已使用分层或迭代方法通过将复杂问题分解为多个子问题来处理长文本。然而,这些方法无法捕获子问题之间的关系。

3 方法

本文提出的自控记忆 (SCM) 系统旨在赋予大规模语言模型 (LLM) 存储长期记忆的能力,使其能够处理冗长的输入并在与用户进行多次交互后保留信息。

3.1 系统总览

在本节中,我们介绍了我们提出的 SCM 系统的工作流程。如图 2 所示,我们的 SCM 系统包括三个模块,包括语言模型代理、内存流和内存控制器。这三个模块协同工作以处理冗长的文档并提供更准确和连贯的响应。我们的系统工作流程包括六个明确的步骤,如下所示:

1. 输入获取: 代理通过直接输入或来自外部源依次接收观察 T (即超长文档输入或用户问题)。

2. 记忆激活: 基于当前观察,记忆控制器确定是否有必要为当前用户输入激活记忆。在保证内存激活的情况下,通过执行步骤 3 和 4 检索相关内存 (有关详细内存信息,请参阅 § 3.2 部分)。否则,该过程直接转到步骤 5。§ 3.3.1 提供了一个全面解释内存控制器的控制流程。

3. 记忆检索: 在这一步中,我们利用观察作为查询来识别相关记忆。每个记忆的分数排名是通过考虑两个维度来计算的: 相关性和新近度。关于相关性,我们评估记忆内容与观察的相似程度。关于新近度,我们考虑自上次访问内存以来经过的时间。随后,我们保留了排名靠前的 K 级记忆。

4. 内存重组: 在这一步中,控制器将决定是直接使用原始内存还是汇总内存。如果选择汇总记忆,原始记忆将被压缩。§ 3.3.2 提供了状态压缩过程的详细解释。然后,系统会将检索到的记忆结构化地组合起来,作为此时响应生成的背景信息。

5. Input Fusion: 在这一步中,我们精心设计了一个提示,将重构后的记忆与现在的观察融合起来,作为模型的输入。§ 3.4 中给出了详尽的描述。6. Response Generation: 模型根据前一步的结果生成响应,并将当前的交互,包括观察和响应,合并到记忆流中。详情请参考 §

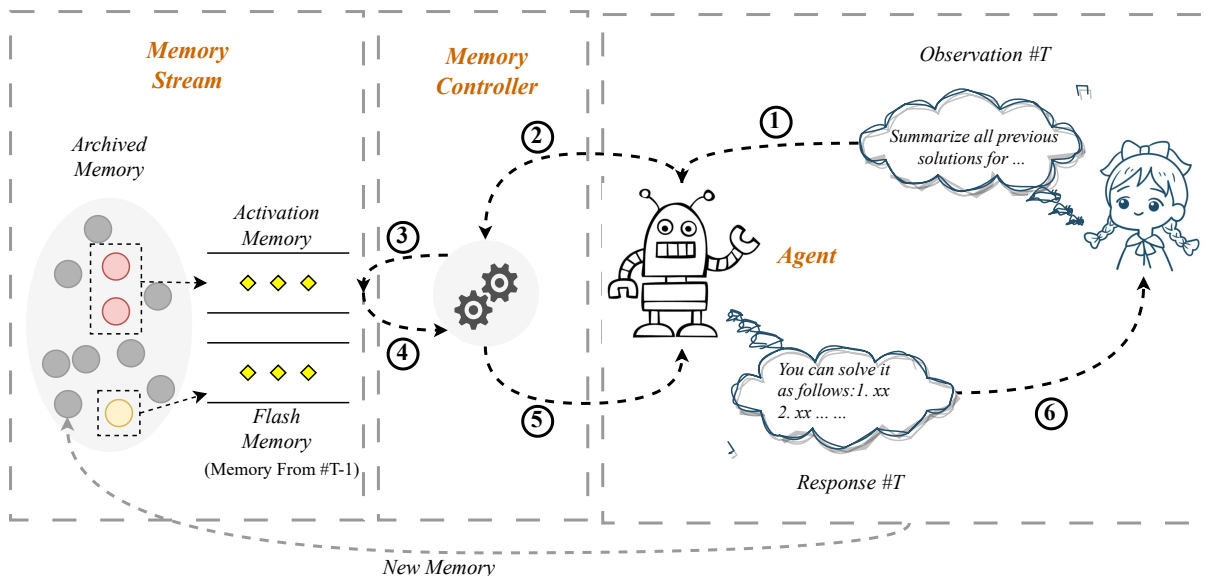


Figure 2: 我们提出的自控记忆 (SCM) 系统的工作流程，其中数字 1-6 表示使用新观测值 # T 进行一次迭代的顺序过程。

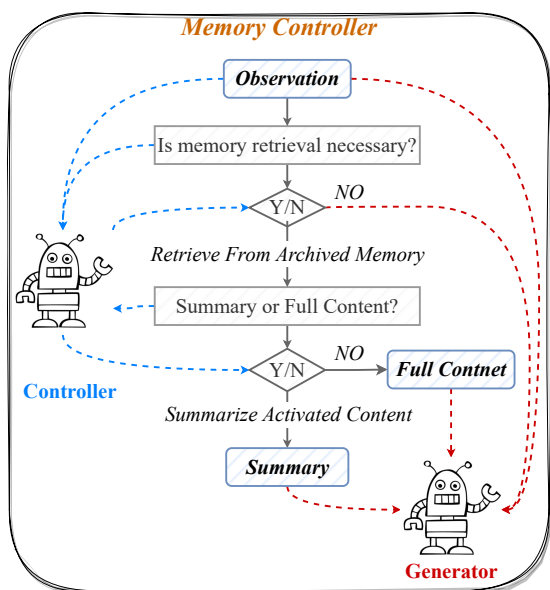


Figure 3: 内存控制器的工作流程。

3.4。

3.2 内存流

本节概述内存流的内部结构。内存流将所有历史内存项存储在一个指定位置，命名为归档内存中心，通过缓存存储和访问工具如 Redis 或 Pinecone⁴ 可以轻松实现高速访问。每个记忆项都包含一个交互索引、一个观察、一个系统响应和一个说明当前交互语义的交互嵌入。此外，The Activation Memory 存储检索到的内存集，Flash Memory 表示 Turn $T - 1$ 的内存。

⁴ 松果：<https://www.pinecone.io/>

3.3 内存控制器

本节讨论使用内存控制器的原因及其工作流程，如图 3 所示。存在三个根本原因。首先，并非所有观察（也称为用户输入或指令）都需要访问历史内存使用情况。例如，用户指令“给我讲个笑话”不需要检索用户的历史记忆。然而，某些用户输入，例如“你还记得我们上周关于健身饮食的结论吗”需要检索过去的记忆。第二个原因是内存量可以很大，从几百到几千甚至几万不等。需要一个控制器来检索和过滤内存。第三个原因是模型的输入长度有限，需要一个 controller 来选择是使用记忆的全文还是记忆的摘要，因为原文可以很长，可能会超过最大长度该模型。接下来的两个小节分别介绍了控制器工作流程和状态压缩实现的细节。

3.3.1 内存控制器工作流程

controller 在过程控制方面的核心是向 agent 提出两个问题：

1. 执行用户命令时是否需要使用内存才能准确回答？
2. 仅使用内存摘要是否可以正常执行用户命令？

第一题提示如图 4，第二题提示如图 5。提示的其他语言版本可以在 § A.1 中找到。

如果控制器确定有必要利用历史记忆，则应进行记忆检索。在检索记忆时，我们使用当前观察（即用户指令）作为查询，并根据两

Given a user command, determine whether executing the command requires historical or previous information, or whether it requires recalling the conversation content. Simply answer yes (A) or no (B) without explaining the information:

Command: [User Input]

Figure 4: 是否需要使用记忆的英文提示。

Given a user command, determine if it can be executed correctly based solely on the summary historical information provided. Simply answer yes (A) or no (B), without explaining the information.

Command: [User Input]

Figure 5: 是否使用摘要记忆的英文提示。

个因素评估每个记忆的排名分数：新近度和相关性。**Recency** 高度优先考虑最近访问的记忆项目，强化了代理的注意力保持在最新交互状态的想法。相关因子为与当前观察相关的记忆项目分配更高的分数。在我们的实现中，我们通过使用语言模型⁵为每个记忆的文本描述创建了一个嵌入向量。计算记忆的嵌入向量与查询观察的嵌入向量之间的余弦相似度以确定相关性。每个记忆的排名分数是其新近度和相关性分数的总和： $rank_score = recency_score + relevance_score$ 。根据长度限制，我们将排名得分最高的顶级 k 记忆指定为激活记忆，其中 k 在 3 到 10 之间变化。

如果控制器确定采用摘要可以允许指令的正常执行，则需要对当前提取的存储器进行摘要。有关内存汇总的具体信息位于以下小节中。

3.3.2 记忆总结

记忆摘要是文档摘要等场景中的一个重要方面，其中单个交互或对话回合的令牌长度可以超过 3000。它可以将多个记忆堆叠到一个激活的记忆部分。图 6 是专门为个人交互（即对话任务）中的记忆总结而设计的英文提示。此外，其他语言版本的提示可以在 § A.3 中找到。

3.4 响应生成

如介绍中所述，SCM 允许非对话优化的 LLM 模拟 ChatGPT。这需要提供明确的指令来提示 LLM 使用激活的闪存来生成所需的响应。图 7 显示了用于超长多交互对话的英文提示，其他语言版本包含在 § A.2 中。

⁵我们使用了 OpenAI 嵌入模型 text-embedding-ada-002

Below is a conversation between a user and an AI assistant. Please provide a summary of the user's question and the assistant's response in one sentence each, with separate paragraphs, while preserving key information as much as possible.

Conversation:

User: [user input]

Assistant: [system response]

Summary:

Figure 6: 提示对话记忆摘要。

Here is a conversation between a user and an AI assistant. Please answer the user's current question based on the history of the conversation:

History of the conversation:

[History of Related Turn]

Previous conversation:

User: [previous user input]

Assistant: [previous system response]

###

User: [current user input]

Assistant:

Figure 7: 超长对话生成英文提示。

4 实验

4.1 设置

我们的框架在两种场景下进行了初步评估：超长对话和超长文档摘要。我们进行实验以回答三个研究问题 (RQ)。

- RQ1. SCM 系统能否在特定令牌限制内与 ChatGPT 竞争甚至超越？
- RQ2. SCM 系统能否扩展以准确回答用户的问题，这些问题与数百甚至数千年前的历史背景相关？
- RQ3. SCM 能否展示对其他场景的泛化，包括长文档摘要？

以下实验评估了与 ChatGPT-Turbo 模型相比，没有对话优化的 text-davinci-003 模型的性



Figure 8: 超长对话示例。

能。评估侧重于超长对话和超长文档摘要的任务。

4.2 定性研究

RQ1. Can SCM system compete with or even outperform ChatGPT within a specific token limit? Yes.

图 1 中的示例包括 4000 个令牌，其中用户询问了他们的爱好，与代理讨论了 100+ 轮。SCM 系统提供了对查询的准确响应，展示了卓越的内存增强功能，从观察中可以明显看出。相比之下，ChatGPT 似乎被大量不相关的历史数据分散了注意力。

RQ2. Can SCM system scale to provide accurate responses to users' questions, which are related to historical contexts that date back hundreds or even thousands of turns? Yes.

图 8 中的示例说明了包含 500 多个回合的超长对话。一开始，用户声明他的目标是减轻体重并打算开始跑步。随后，用户和模型每天就实现减肥目标的进展以及其他对话主题进行交谈。十天后，对话长度达到 10,000 个令牌。然后用户询问模型“你还记得我的冷杉吗”。我们的 SCM 系统准确地回答了这个问题。

RQ3. Can SCM demonstrate generalization to other scenarios, including long document summarization? Yes.

图 9 说明了一个令人难以置信的冗长文档摘要的实例。具体来说，该报告涉及 OpenAI 对 GPT-4 的揭幕。超过 4,000 个字符的摘要对传统模型提出了挑战，因此需要对文档部分进行拆分和单独摘要，然后将其合并。尽管如此，这种方法会失去段落之间的依赖关系。我们的框架使用迭代总结过程。在总结段落时，我们的方法依赖于早期相关的摘要记忆来生成更精确的摘要。最终，该框架结合了分而治之的策略来生成最终的文档摘要。通过分而治之的方法获得的最终摘要利用来自每个文档块的信息提供了一个全面的摘要。此外，我们的迭代摘要范例包含一个记忆增强功能，允许通过在单块摘要期间集成问题提问方法来生成特定主题的摘要。例如，如果用户提出诸如“请提供 GPT4 在图像处理方面的技术细节和评估结果的摘要”之类的问题，该模型将访问先前的摘要记忆并提取相关内容。未来我们会继续改进这方面的内容。

5 限制和风险

限制 缺乏适当的数据集来评估对极长文本的处理，导致我们的模型只能通过手动验证来验证。然而，这种方法不足以全面客观地评估不同的情景。因此，我们的目标是构建一个特

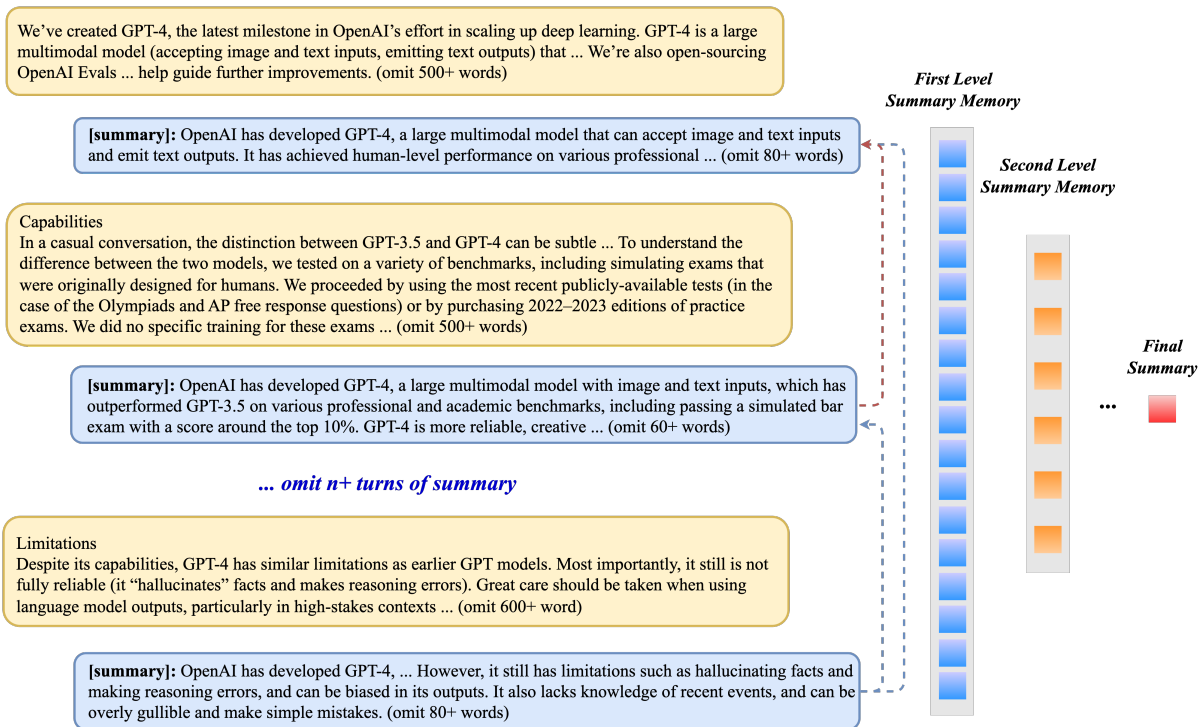


Figure 9: 超长迭代和分层摘要示例。

定的测试集，其中包含在不同环境中处理长文本所必需的各种关键指标。该测试集将附有手动评估标准，以便与相关方法进行更公平的比较。此外，我们将评估我们的系统在更多具有单轮指令理解能力的开源模型上的功效。

风险 我们的系统能够附加到任何 LLM，这可能容易出现事实错误、妄想、有毒语言和恶意响应。因此，我们目前将系统的使用限制在学术研究目的。

6 结论和未来的工作

在本文中，我们提出了一个自控记忆 (SCM) 系统，将任何 LLM 模型的输入长度扩展到无限长度，并有效地从所有历史信息中捕获有用的信息。该方法不需要对模型进行任何训练和修改，具有很强的适用性。我们通过手动评估 ChatGPT 和基于我们系统的 Text-DaVinci-003 模型验证了我们方法的有效性，与 ChatGPT 相比，在长文本场景的某些方面展示了优越的性能。

我们未来的工作将集中在发布一个全面的测试集及其手动评估标准，并在当前可用的各种开源模型上测试我们的系统。

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain,

Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. [Training a helpful and harmless assistant with reinforcement learning from human feedback.](#)

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. [Longformer: The long-document transformer.](#)

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. [Language models are few-shot learners.](#) In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child,

- Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. [Language models are few-shot learners.](#)
- Shuyang Cao and Lu Wang. 2022. [HIBRIDS: Attention with hierarchical biases for structure-aware long document summarization.](#) In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 786–807, Dublin, Ireland. Association for Computational Linguistics.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. [Palm: Scaling language modeling with pathways.](#)
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models.](#)
- Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2023. [A survey of natural language generation.](#) *ACM Comput. Surv.*, 55(8):173:1–173:38.
- Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. [LongT5: Efficient text-to-text transformer for long sequences.](#) In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. [Crosslingual generalization through multitask finetuning.](#)
- OpenAI. 2022. [Introducing chatgpt.](#)
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback.](#)
- Jason Phang, Yao Zhao, and Peter J. Liu. 2022. [Investigating efficiently extending transformers for long input summarization.](#)
- Ofir Press, Noah Smith, and Mike Lewis. 2022. [Train short, test long: Attention with linear biases enables input length extrapolation.](#) In *International Conference on Learning Representations.*
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. [Improving language understanding with unsupervised learning.](#)
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners.](#)
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer.](#) *Journal of Machine Learning Research*, 21(140):1–67.
- Xiaozhe Ren, Pingyi Zhou, Xinfan Meng, Xinjing Huang, Yadao Wang, Weichao Wang, Pengfei Li, Xiaoda Zhang, Alexander Podolskiy, Grigory Arshinov, Andrey Bout, Irina Piontkovskaya, Jiansheng Wei, Xin Jiang, Teng Su, Qun Liu, and Jun Yao. 2023. [Pangu- \$\Sigma\$: Towards trillion parameter language model with sparse heterogeneous computing.](#)
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms.](#)
- Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. [Learning to summarize from human feedback.](#) *CoRR*, abs/2009.01325.
- Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic.

2022. [Galactica: A large language model for science](#).
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, and etc. 2022. [Lamda: Language models for dialog applications](#).
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. [Linformer: Self-attention with linear complexity](#).
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. [Finetuned language models are zero-shot learners](#). In *International Conference on Learning Representations*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. [Emergent abilities of large language models](#). *Transactions on Machine Learning Research*. Survey Certification.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022c. [Chain of thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*.
- BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ili, Daniel Hesslow, Roman Castagné, Alexandra Sasha Lucioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, and etc. 2023. [Bloom: A 176b-parameter open-access multilingual language model](#).
- Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. [Recursively summarizing books with human feedback](#).
- Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. [Big bird: Transformers for longer sequences](#).
- Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. [GLM-130b: An open bilingual pre-trained model](#). In *The Eleventh International Conference on Learning Representations (ICLR)*.
- Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, Chen Li, Ziyang Gong, Yifan Yao, Xinjing Huang, Jun Wang, Jianfeng Yu, Qi Guo, Yue Yu, Yan Zhang, Jin Wang, Hengtao Tao, Dasen Yan, Zexuan Yi, Fang Peng, Fangqing Jiang, Han Zhang, Lingfeng Deng, Yehong Zhang, Zhe Lin, Chao Zhang, Shaojie Zhang, Mingyue Guo, Shanzhi Gu, Gaojun Fan, Yaowei Wang, Xuefeng Jin, Qun Liu, and Yonghong Tian. 2021. [Pangu- \$\alpha\$: Large-scale autoregressive pretrained chinese language models with auto-parallel computation](#).
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022a. [Opt: Open pre-trained transformer language models](#).
- Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022b. [Summⁿ: A multi-stage summarization framework for long input dialogues and documents](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics.

A 提示列表

A.1 内存控制器提示

给定一个用户指令，判断执行该指令是否需要历史信息或者上文的信息，或者需要回忆对话内容，只需要回答是(A)或者否(B)，不需要解释信息：

指令：[用户输入]

Figure 10: 内存控制器中文提示。

A.2 对话生成提示

A.3 提示对话状态压缩

以下是用户和人工智能助手的对话，请根据历史对话内容，回答用户当前问题：

相关历史对话：

[历史轮对话内容]

上一轮对话：

[上一轮对话内容]

###

用户：[用户问题]

助手：

Figure 11: 超长对话生成中文提示。

以下是用户和人工智能助手的一段对话，请分别用一句话写出用户摘要、助手摘要，分段列出，要求尽可能保留用户问题和助手回答的关键信息。

对话内容：

用户：[用户输入]

助手：[系统回复]

摘要：

Figure 12: 超长对话摘要中文提示。