arXiv:2305.04782v2 [cs.CL] 3 Dec 2023

HISTALIGN：通过与历史记录对齐，提升语言生成中的上下文依赖

David Wan ShiyueZhang Mohit Bansal北卡罗来纳大学
教堂山分校 {davidwan, shiyue, mbansal}@cs.unc.edu

摘要

语言模型（语言模型）可以生成幻觉和不连贯的输出，这凸显了其较弱的上下文依赖。缓存语言模型通过利用近期历史记录的记忆来增强语言模型，可以提高上下文依赖，并且在多种语言生成任务中展现出卓越性能。然而，我们发现，即使经过训练，当前缓存语言模型的缓存组件所带来的性能提升仍不理想，这是由于当前隐藏状态与记忆中存储的隐藏状态之间存在错位。在本研究中，我们提出了 HISTALIGN，这是一种新的训练方法，旨在确保良好的缓存对齐，使模型能够从历史记录中获得有用信号。我们首先在一个简单的合成任务上验证了这一概念，在该任务中，记忆对于正确预测至关重要；结果表明，HISTALIGN 的缓存组件具有更好的对齐效果，并提升了整体性能。接下来，我们在多种下游语言生成任务上评估 HISTALIGN，包括提示续写、抽象摘要和数据到文本。我们证明，在开放式生成和条件生成设置下，HISTALIGN 分别提升了文本连贯性和忠实度。HISTALIGN 还可推广至不同的模型系列，展现出其在多种场景下提升语言模型上下文依赖的能力。[1]

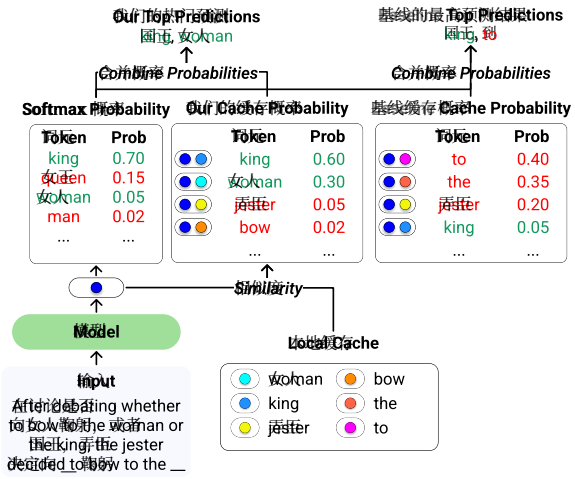


图 1：HISTALIGN 与基线缓存语言模型的示意图。输入示例摘自Chang 和 Mc- Callum（2022年）。我们的 HISTALIGN 能够为国王 和女人赋予较高的概率，因此能够降低由 softmax 概率赋予幻觉词元女王的权重。当前的缓存语言模型（基线）会为缓存中的无关词元赋予较高概率，因此有可能生成幻觉或不连贯的词元。

1 引言

语言建模（LM），也称语言生成，需要具备良好的上下文依赖性。对于开放式和条件生成任务，我们希望模型生成的内容与此前生成的内容或输入上下文保持一致。然而，当前模型生成中普遍存在不连贯和幻觉问题（Holtzman 等人，2020年；Cao 等人，2018年；Maynez 等人，2020年），这表明语言模型的上下文依赖性较弱。

缓存语言模型（Grave 等人，2017年b，缓存语言模型）是一种简单而有效的方法，它通过为语言模型配备额外的近期历史记录记忆（局部上下文），并使其能够直接从历史记录中“复制”，从而增强上下文依赖性。这类模型在语言建模和下游生成任务中取得了显著提升（Merity 等人，2017年；See 等人，2017年）。然而，自 Transformer 问世以来（Vaswani 等人，2017年），得益于强大的自注意力机制，局部记忆的使用减少了，越来越多的研究开始关注如何利用长期记忆或外部记忆（Khandel- wal 等人，2020年；Yogatama 等人，2021年）。尽管如

[1]我们的代码已公开发布于 https://github.com/meetdavidwan/histalign

此，Zhong 等人（2022年）表明，在 Transformer 上使用局部记忆仍然有益。

本文重点研究如何将本地缓存应用于基于 Transformer 的语言模型，并表明，缓存组件越能与模型对齐，带来的增益就越大。首先，我们从理论上证明，缓存语言模型能够打破限制参数化语言模型对高度上下文相关的自然语言进行建模能力的 softmax 瓶颈（Yang 等，2018年）。随后我们发现，在当前的缓存语言模型中，即使在训练期间使用缓存组件，记忆组件所提供的信号也很微弱（Zhong 等，2022年）。我们假设，主要瓶颈源于当前隐藏状态与记忆中的隐藏状态之间的错位，因此，相关性更高的记忆并未获得比相关性较低的记忆更高的权重。我们通过一个合成任务“歧义模板”（Chang 和 McCallum，2022年）展示了这一问题，其示例见图 1。当要求模型根据上下文“在犹豫是向女人还是国王鞠躬之后，弄臣决定向__ 鞠躬”预测下一个词时，当前的缓存语言模型并未将期望的词国王和女人赋予最高概率。相反，我们发现，无关词（例如向和弄臣）具有较高的缓存概率。将这些概率与原始 softmax 概率结合时，期望的词无法排在词元首位。我们发现，这一问题存在于不同规模的预训练语言模型、微调模型以及经过缓存增强的模型中。

接下来，我们提出一种新的微调方案 HISTALIGN，以解决这一错位问题。该方案通过在语言模型训练目标中加入对比损失，鼓励模型将当前隐藏状态与历史记录中的隐藏状态对齐。如图 1所示，我们的缓存组件为国王和女人赋予的概率高于缓存中其他相关性较低的词。不同于将所有负例一视同仁的典型对比损失，我们提出学习负词元的排序，即语义上越相似的词元，排名越高。如图 2所示，当我们对齐住房对应的空间时，希望住宿 等词比儿童等相关性较低的词更接近。因此，即使历史记录中没有出现完全匹配的目标词，缓存仍然有用。我们通过合成歧义模板任务展示 HISTALIGN 更强的缓存性能，并展示它在提升开放式提示续写的连贯性，以及提高抽象式摘要和数据到文本任务的忠实度方面的优势。

总而言之，我们的贡献如下：

- 我们从 softmax 瓶颈的角度讨论了为何采用局部记忆的缓存语言模型能够提升上下文依赖。
- 我们揭示了当前缓存语言模型及其训练策略中存在的错位问题。
- 我们提出了一种基于顺序信息对比学习的新训练方法 HISTALIGN，以缓解错位问题并更好地利用记忆。
- 我们证明，HISTALIGN 能够提升开放式生成的连贯性以及条件生成的忠实度，并且适用于不同的模型系列，计算开销也很小。

2 相关工作

缓存语言模型和指针网络。向语言模型（LM）中添加缓存组件最早由 Kuhn 和 De Mori（1990年）为语音识别提出。Grave 等人（2017年c）将这一思想扩展到基于 RNN 的神经语言模型，并将其称为神经缓存语言模型。缓存语言模型通过结合 RNN 模型的输出以及缓存与当前隐藏状态之间的相似度来预测下一个词元。缓存会保存近期历史记录中的隐藏状态和下一个词元预测组成的元组，即（h_i，x_{i+1}）（见第 3.2 节）。从本质上说，缓存组件使模型能够从历史记录中复制词元。与缓存语言模型类似，指针网络（Vinyals 等人，2015年；Merity 等人，2017年）也结合了词元的生成与复制，但使用 h_i 作为 x_i 的表示（而非 x_{i+1}）。这意味着，指针网络需要学习当前表示与过去表示之间的额外变换，以及用于插值的门控组件（Merity 等人，2017年；See 等人，2017年）。[2] 相比之下，缓存语言模型无需学习额外参数，可以直接在测试时应用。对于更大的缓存大小而言，它的效率更高，

[2]根据具体实现，模型还可能包含与隐藏维度的数量呈二次关系的额外参数，用于投影矩阵（例如，See 等人 (2017年) 在门控模块中使用了四个隐藏状态的拼接）。

可用于更大的缓存大小（即将缓存语言模型扩展到长期记忆和外部记忆），并且已被证明其性能优于指针网络（Grave 等人，2017年b；Zhong 等人，2022年）。

虽然缓存语言模型可以在测试时直接应用，但一项近期工作（Zhong 等，2022年）表明，在训练时也使用缓存能够带来更大的提升。不过，这类针对缓存语言模型提出的学习目标通常只为缓存组件提供远程监督。相比之下，我们对缓存引入了直接监督，使当前表示与其历史记录保持一致。

具有局部或外部记忆的语言模型。缓存语言模型和指针网络最初仅使用局部上下文中的隐藏状态，即输入上下文中的前置词元。尽管已有研究证明，这项技术有助于语言建模和其他语言生成任务（Gulcehre 等人，2016年；Grave 等人，2017年c；Merity 等人，2017年；See 等人，2017年），但在 Transformer架构流行后，它的使用有所减少，因为自注意力机制能够关注输入上下文中的任意词元。因此，许多工作（Grave 等人，2017年a；Khandelwal 等人，2020年；Yogatama 等人，2021年；Zhong 等，2022；Min 等人，2022年）提出通过应用检索技术，使用超越局部上下文的长期记忆或外部记忆。尽管我们的方法可以扩展到外部缓存设置，但我们仅关注融入局部记忆，并表明局部记忆在 Transformer 之上仍然有益，因为它打破了参数化语言模型的 softmax 瓶颈（Yang 等人，2018年）。一项同期工作（Chang 等人，2023年）也通过示例和实证结果展示了指针网络如何打破 softmax 瓶颈，而我们则在第 4.1 节中以更具数学性的方式对此进行讨论。

语言生成中的上下文依赖。现有的语言生成模型表现出较弱的上下文依赖。对于开放式生成任务，Holtzman 等人(2020年) 指出，强大的语言模型在输入提示之后可能会生成非常不连贯的文本。故事生成领域也长期观察到这一不连贯问题（Rashkin 等人，2020年；Alabdulkarim 等人，2021年）。对于条件生成任务，例如摘要，Cao 等人(2018年)；Maynez 等人(2020年) 表明，约 30% 和 70% 模型生成的摘要在两个广泛使用的摘要数据集上分别包含幻觉。数据到文本生成（Chen 等人，2020a）、机器翻译（Weng 等人，2020年）等任务中也出现过类似的不忠实性问题。尽管已有许多方法被提出，用于缓解不连贯性（Li 等人，2022a）或不忠实性问题（Cao 和 Wang，2021年；Wan 和 Bansal，2022年），本文探索一种简单而通用的缓存语言模型方法，以提升多种任务对上下文的依赖程度。同时开展的工作（Chang 等人，2023年）采用指针网络类型的架构来改进下一个词的概率分布和摘要事实性。他们通过使用额外的上下文相关嵌入来修改 softmax 头。相比之下，我们只应用原始的缓存语言模型架构，并通过一种新颖的训练目标对其进行改进。

3 预备知识

3.1 语言建模

<escape>我们专注于自回归语言建模（语言模型）。在这里，为了简化起见，我们假设语言模型是仅解码器架构，即当前步骤的上下文是前面步骤生成的词元。我们将在第 4.3 节展示，同样的方法可以轻松推广到编码器-解码器模型。给定上下文 $c_t = x_1$, ……, x_{t-1}，下一个词元 $x_t = w$ 的概率由一个 softmax 头预测：</escape>

$$P_{lm}(w|c_t) \propto \exp(h_t^\top e_w) \tag{1}$$

其中， e_w 是词元 w的输出嵌入， h_t 是模型在第 t步的输出上下文向量（隐藏状态）。模型通过最小化交叉熵损失进行训练： $l_{xe} = -\sum_t \log P_{lm}(x_t|c_t)$。

3.2 缓存语言模型

缓存语言模型为语言模型增添了一个记忆组件。遵循Grave 等人(2017年c)，我们将缓存视为由上下文向量和目标词元组成的元组列表，(h_i, x_i)。假设我们只考虑局部上下文的历史记录，那么 t步的局部记忆写作：

$$\mathcal{M}_{\text{local}} = \{(h_i, x_i)\}_{1 \le i \le t-1} \tag{2}$$

然后，下一词元预测会汇总 softmax 头输出的对数值，以及与记忆中保存的

h_t 之间的相似度：

$$P_{clm}(w|c_t) \propto \exp(h_t^\top e_w) + \sum_{(h_i,x_i)\in\mathcal{M}_{\text{local}}} \mathbb{1}_{\{x_i=w\}} \exp(\text{sim}(h_t, h_i)) \quad (3)$$

其中 sim$(\cdot,\ \cdot)$ 可以是任意相似度函数。这里，我们遵循 Zhong 等人(2022年) 并使用缩放点积：sim$(h_1,\ h_2) = \frac{h_1 \cdot h_2}{\sqrt{d}}$，其中$d$是隐藏维度大小。

虽然 Grave 等人 (2017c) 仅在评估时使用缓存，但 TRIME (Zhong 等人，2022年) 表明，在训练期间也使用缓存能带来更多益处，即最小化$l_{trime} = -\sum_t \log\ P_{clm}(x_t|c_t)$。这里，我们在训练和评估期间都使用缓存，但通过对缓存引入直接监督来改进训练目标（见第 4.2节）。

4 我们的方法

4.1 突破 softmax 瓶颈

我们首先希望将使用局部记忆与 softmax 瓶颈问题联系起来（Yang 等人，2018年），并表明 Transformer 的自注意力无法突破这一瓶颈，而局部缓存可以。

参数化自回归语言模型（第 3.1 节），包括基于 Transformer 的语言模型，会对上下文向量（或隐藏状态）H $\in \mathbb{R}^{N\times d}$ 与输出嵌入矩阵 E $\in \mathbb{R}^{V\times d}$进行运算，并使用 softmax 函数。 N 是上下文的数量；假设训练集中的每个词元都有不同的上下文，那么 N 就是训练集中的词元数量。 V是词汇表大小， d 是隐藏维度大小。于是，下一个词元的概率构成一个对数概率矩阵 A $\in \mathbb{R}^{N\times V}$($A_{tw} = \log\ P(w|h_t)$)。理想情况下，由于每个上下文都是唯一的，A 的秩应尽可能大，达到 V（假设 $V < N$）。然而，由于 A 大致等价于 HE$^\top$，其秩严格受限于隐藏维度 d （形式化证明请参阅 Yang 等人（2018））。这一低秩问题极大地限制了语言模型对高度依赖上下文的自然语言进行建模的能力。如图 1 所示，queen 的概率高于 woman。Chang 和 McCallum（2022）对语言模型难以处理这种双峰分布的解释是：king、woman、man、queen 这四个词往往会在嵌入空间中构成一个平行四边形；如果模型的隐藏状态希望靠近 king 和 woman 的输出嵌入，它也会靠近 man 和 queen 的输出嵌入。

要突破这一瓶颈，一个简单的解决方案是增大 d，因为我们知道，模型通常越大，性能越好。Yang 等人（2018）提出、并由 Kanai 等人（2018）、Yang 等人（2019）以及 Chang 和 McCallum（2022）拓展的另一种方案，是使用多个 softmax 头——混合 softmax（MoS），例如 $P(w|h_t)$ $\propto$exp($h_t^{(1)\top} e_w$) +exp($h_t^{(2)\top} e_w$) + exp($h_t^{(3)\top} e_w$)。每个 $h_t^{(k)}$ 都是不同的上下文向量。然而，增加 softmax 头的计算开销相当大。将 MoS 与公式 3 对比可见，添加 exp($h_t^\top e_w$) 和 exp(sim(h_t, h_i))，就类似于 MoS，但不需要增加额外的 softmax 头。理解这种联系的另一种方式是：使用局部记忆时，A 大致等价于 HE$^\top$ + HH$_c^\top$，其中 H$_c$ 是局部上下文中的隐藏状态。[3] 假设 E$_c$ = E +H$_c$，则 A 变为 HE$_c$。与 E 不同，E$_c$ 不再是大小为 $V \times d$的静态输出嵌入矩阵，而是大小为$N \times V \times d$的上下文相关嵌入张量。因此，A 的秩不再受限于 d。请注意，这种联系同样适用于使用长期记忆或外部记忆的情况。

4.2 HISTALIGN

缓存语言模型通过聚合当前隐藏状态与缓存中隐藏状态之间的相似度分数，将原始 softmax 概率与缓存概率结合起来。要有效使用缓存模块，相似度函数 sim$(\cdot,\cdot)$ 在公式 3中发挥着重要作用。如果当前隐藏状态与相关性较低的记忆之间的相似度高于与相关性较高的记忆之间的相似度，就会使模型偏离从缓存中选择最有用信息的方向。通过为正确的局部记忆（例如，在国王 和 女人的示例中，如图 1 所示）赋予较高的概率，我们可以确保在合并概率时，这些记忆的得分高于无关及幻觉生成的词元。然而，我们发现，即使直接最大化 log P_{clm} (Zhong 等人, 2022年)，也无法保证当前表示与记忆中存储的相关信息良好对齐，图 1 中的基线概率便展示了这一点（更多详情请参见第 6.1 节）。

因此，为了应对这种错位，我们提

[3]这是因为，在缓存语言模型设置下，logits 是原始词元logits $HE^\top$ （公式3的第一项）与缓存logits $HH_C^\top$ （公式3的第二项）之和，而Yang 等人（2018年）的性质 2 表明，对数概率矩阵与logit矩阵的秩相近。

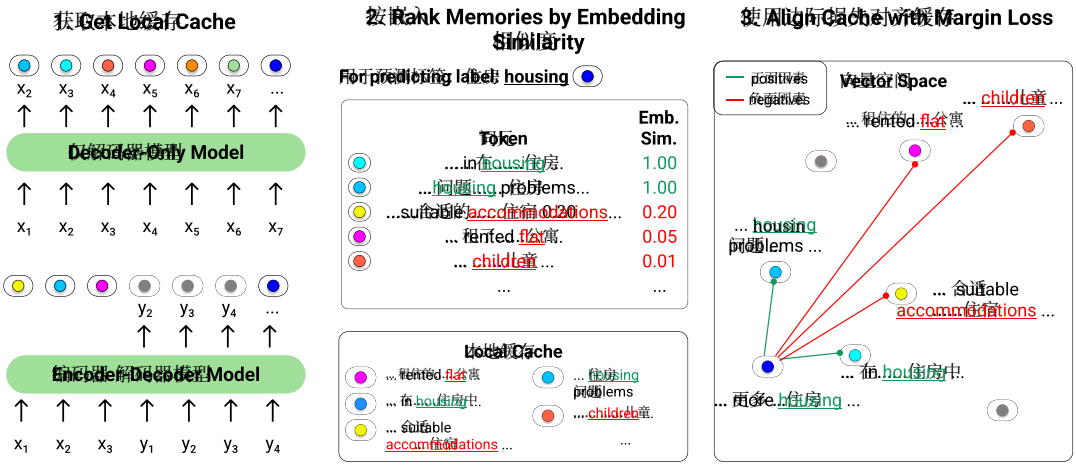


图 2：HISTALIGN 训练方法示意图。我们首先将局部上下文中的隐藏状态与对应的目标词元结合，得到本地缓存，然后根据嵌入相似度对其排序。接着使用排序后的记忆，通过间隔损失进行训练。这样可以确保相似但为负例的词语（例如住宿）在向量空间中比无关词语（例如儿童）更接近。

出一种新的对比目标，鼓励相似目标词元的隐藏状态具有更高的相似度。在训练过程中，给定当前隐藏状态 h_t 和对应的下一个词元x_t，我们通过从缓存中选择具有相同目标词元的记忆来构建正样本集 $\mathcal{P}_t$：

$$\mathcal{P}_t = \{(h_i, x_i)\}_{x_i = x_t, 1 \le i \le t-1} \tag{4}$$

其他所有记忆均视为负例。步骤 2 中展示了一个示例，见图 2。在预测词元住房时，我们有两处此前提及“住房”的记录，而其他词语，包括公寓、儿童、住宿等，都被视为负例。

在典型的对比损失中，例如 InfoNCE（van den Oord 等人，2019年），所有负例都一视同仁。然而，我们希望学习对负例进行排序：相似度较高的示例排在相似度较低的示例之前。在图 2的示例中，住宿比住房更相似于儿童。这样一来，即使要预测的词语在本地缓存中没有此前的提及，我们的模型仍能输出合理的替代词。

为实现这一点，我们通过计算当前目标词的嵌入与缓存中各词嵌入之间的余弦相似度，构建记忆的排序，即 cosim(e_t, e_i)。将词元按语义相似度从高到低排序后，我们使用以下最大间隔损失（Liu 等人，2022年c）：

$$l_{cont.} = \sum_t \sum_{i \in \mathcal{P}_t} \sum_{j > i, j \notin \mathcal{P}_t} \max\left(0, \mathrm{sim}(h_t, h_j) - \mathrm{sim}(h_t, h_i) + \lambda_{i,j}\right) \tag{5}$$

其中 $\lambda_{i,\ j} = (j - i)\lambda$，而 λ 是根据验证损失调优的间隔。

HISTALIGN 的最终目标是原始语言模型交叉熵损失 l_{xe} 与基于排序的对比损失的组合：

$$l_{histalign} = l_{xe} + \alpha l_{cont.} \tag{6}$$

其中 α 是对比损失的可调权重。请注意，在推理时，我们使用公式 3。

4.3 扩展到编码器-解码器模型

HISTALIGN 可以轻松适配编码器-解码器模型。对于条件生成任务，目标文本通常较短，因此连贯性并非大问题。更关键的是，生成的目标内容是否忠实于输入上下文，例如摘要任务中的输入文档，或数据到文本任务中的输入表格。因此，我们将输入词元及其对应的编码器隐藏状态定义为本地缓存，而仅解码器模型的本地缓存则包含输出词元和解码器隐藏状态。然后，我们计算当前解码器隐藏状态与缓存中存储的编码器隐藏状态之间的相似度。

5 实验设置

这里，我们介绍任务和实验设置。详情请参阅附录A 。

5.1 任务与数据集

歧义模板 是由 Chang 和 McCallum (2022年整理的一种实用合成数据集，其中的每个示例都使用取自谷歌（英语）类比数据集（[4] Mikolov 等人，2013）中语义类比关系的对角词构造模板生成。这种简单而有效的设置可用于检验模型能否从历史记录中复制正确的词元，而不会臆造语义相似的词元，例如女王和 男人 这两个词在图 1所示示例中的用法。由于目标词总能在上下文中找到，我们还可以仅使用缓存组件来评估性能。

开放式生成 通过要求模型根据提示词生成后续内容，来评估其语言建模能力（Holtzman 等人，2020年；Su 等人，2022年；Li 等人，2022b）。我们使用 WritingPrompts（Fan 等人，2018年），将前 50 个词元作为提示词，并允许模型使用经典的核采样生成最多 256 个词元（$p = 0.95$）（Holtzman 等人，2020年）。

抽象式摘要 是指提供输入文档的精简版本这一任务。一个关键问题是“幻觉”，即生成摘要中包含错误或文档中不存在的事实或实体（Cao 等人，2018年；Maynez 等人，2020年）。我们在两个广泛使用的英文新闻摘要数据集 XSum（Narayan 等人，2018年）和 CNN/DM（Hermann 等人，2015年）上进行评估。

数据到文本 是描述结构化数据的任务，忠实度极其重要，因为在描述医疗报告或金融统计数据等情形中，人们无法容忍任何幻觉（Thomson 和 Reiter，2020年）。我们在 LogicNLG 上进行评估（Chen 等人，2020a）。

5.2 系统

我们使用 GPT2-small 和 GPT2-large（Radford 等，2019年）进行歧义模板和提示词续写，并使用 BART-large（Lewis 等，2020年）进行摘要和数据到文本任务。对于所有任务，我们选择对预训练语言模型进行微调。

[4]这指的是在嵌入空间中位于平行四边形对角线上的词。例如，对于元组（女王、国王、女人、男人），对角词是（国王、女人）和（女王、男人）。

我们比较的第一个基线是使用原始交叉熵损失进行微调（l_{xe} 见第3.1节），在我们的结果表中以原始模型名称标示。随后，我们还将其与最新的缓存语言模型学习目标 TRIME （Zhong 等，2022年）（l_{trime} 见第3.2节）进行比较。

5.3 评估

歧义模板。 作为一项概念验证实验，我们在完整设置下使用公式中的组合概率进行评估 3，同时也在仅缓存设置下进行评估，仅使用缓存相似度分数来预测下一个词元。我们通过两个对角词出现在前k 个预测中的准确率（Acc@k）来评估性能，其中 $k = \{2, 5, 10, 25\}$。理想情况下，当 $k = 2$时，我们希望准确率达到 100%，这表明两个对角词是排名前 2 的选项。请注意，仅使用缓存时， k值为 50 即可达到完美准确率，因为这会包含整个局部历史记录。此外，我们希望通过实证验证，带有局部记忆的缓存语言模型能够突破 softmax 瓶颈。为此，我们在完整设置下使用 500 个示例（具体而言，对于基于 GPT-2 的模型，使用 $\mathbf{A} \in \mathbb{R}^{N \times V}$（第 4.1 节）中的对数概率矩阵的秩，其中 $N = 4750$和 $V = 50257$ ）。

开放式生成。 我们主要评估连贯性，即模型生成的续写的连贯性。遵循Su 等人 (2022年) 的做法，连贯性通过提示词与续写的 SimCSE（Gao 等人，2021年）句子嵌入的余弦相似度来近似衡量。此外，遵循以往研究，我们报告 n-gram 多样性（Meister 等人，2022年）和 MAUVE (Pillutla 等人，2021年) 分数，以进行更全面的评估。我们希望 HISTALIGN 不会损害多样性和 MAUVE。我们还在 Amazon MTurk 上进行**人工评估**，请工作人员比较 TRIME 和 HISTALIGN生成的续写。更多详情见附录B.1。

抽象式摘要。 我们主要使用三种广泛采用的自动指标来评估生成摘要的忠实度：FactCC（Kryscinski 等人，2020年）和 DAE（Goyal 和 Dur- rett，2021年），它们都是基于蕴含的指标；以及实体精确率（Nan 等人，2021年，P_{ENT}），该指标计算摘要中出现在文档里的实体所占的比例。我们还报告 ROUGE-L（Lin，2004），用于评估一般性的内容选择。同样，我们还进行**人工**

	Full				仅缓存				Full
模型	Acc@2	Acc@5	Acc@10	Acc@25	Acc@2	Acc@5	准确率@10	Acc@25	Rank
GPT2-Small	50.00	62.20	68.80	76.96	0.00	35.57	50.56	75.71	762
TRIME	46.43	56.41	76.47	97.96	0.00	39.89	66.51	**100.00**	836
HISTALIGN	**63.47**	**72.26**	**89.71**	**100.00**	**58.62**	**70.62**	**79.59**	94.31	**854**
GPT2-large	75.43	84.76	87.93	91.26	0.05	37.83	73.88	**100.00**	1280
TRIME	77.40	91.10	94.56	97.17	0.11	22.32	84.89	**100.00**	**1377**
HISTALIGN	**82.22**	**92.84**	**96.57**	**98.34**	**82.15**	**92.51**	**99.94**	**100.00**	**1377**

表 1：AmbiguousTemplate 上的结果。HISTALIGN 在完整设置和仅缓存设置下均取得了最佳性能。我们还通过实验证明，TRIME 和 HISTALIGN 打破了 softmax 瓶颈。

模型	Acc@2	Acc@5	Acc@10	Acc@25
LLaMA2-7B	0	0	0	**100**
TRIME	0	0	0	**100**
HISTALIGN	**100**	**100**	**100**	**100**

表2：LLaMA2-7B模型在歧义模板上的仅缓存结果。

评估，我们请众包工作者判断每篇摘要（从 100 个随机抽取的样例中选出）是否忠实且信息丰富。更多详情请参阅附录B.2。

数据到文本。我们主要通过 NLI-Acc 和 SP-Acc（Chen 等人，2020a）以及另外两个较新的指标——TAPEX-Acc 和 TAPAS-Acc（Liu 等人，2022a）评估模型生成内容的忠实度。NLI-Acc 是一种基于蕴含的指标，使用 TaBERT（Chen 等人，2020b）在 TabFact 数据集（Yin 等人，2020年）上预训练；SP-Acc 则先将句子解析为逻辑程序，再评估执行准确率。TAPEX-Acc 和 TAPAS-Acc 分别是使用 TAPEX（Liu 等人，2022b）和 TAPAS（Eisen-schlos 等人，2020年）训练的基于蕴含的指标。与以往工作（Chen 等人，2020a）相同，我们报告 BLEU（Pap-ineni 等人，2002）作为表层评估指标。

6 结果

我们验证了 HISTALIGN 在对齐缓存组件方面的能力，并因此改善了第 6.1 节歧义模板上的下一个词元预测；第6.2 节开放式提示词续写的连贯性；以及第 6.3 节和第 6.4 节抽象式摘要和数据到文本任务的忠实度。

6.1 缓存对歧义模板的重要性

我们在表 1中展示了歧义模板的结果。首先可以看出，原始 GPT2 模型在仅缓存设置下表现相当糟糕，尤其是 Acc@2。这是预料之中的，因为原始模型在微调时使用了交叉熵损失，且未涉及缓存组件，因此在测试时应用缓存可能并无帮助。其次，尽管 TRIME（Zhong 等人，2022年）在完整设置下的表现通常优于原始模型，但其仅缓存设置下的 Acc@2 和 Acc@5 与原始模型相近。考虑到所有目标词都出现在历史记录中，这一结果表明，尽管 TRIME 在训练期间使用了缓存，其缓存组件仍未对齐，对最终性能的贡献有限。

相比之下，HISTALIGN 仅使用缓存模块便取得了较高的 Acc@2，在两种模型规模下都显著优于原始模型和 TRIME ，这证明了我们的对比损失有助于更好地对齐记忆。因此，HISTALIGN 在完整设置下的所有 k 上均优于两种基线模型。此外，两种模型规模都能获得提升，不过大型模型的差距较小。这一观察结果与我们在第 4.1 节中的讨论一致：隐藏维度更大的模型受 softmax 瓶颈的影响较小，而本地记忆有助于打破任何参数化语言模型的这一瓶颈。表明对数概率矩阵秩的 表 1 也从实证角度验证了这一点：原始模型的秩受其隐藏维度的上限约束（GPT2-small 为 768，GPT2-large 为 1280），而加入本地缓存则打破了这一瓶颈。最后，我们在 表 9 中展示了两个定性示例。详细讨论请见附录 C。

在近期大语言模型上进行实验。 我们还对 LLaMA2 7B 模型进行了微调（Touvron 等人，2023年）。有趣的是，我们发现，零样本评估时，LLaMA2 在 Acc@{2,5,10} 上的准确率为 0%。微调后，该模型无需任何缓存即可达到 100% 的准确率。这是意料之中的，因为该任务是一个简单的合成任务，而且与 GPT2-large 相比，该模型大了 10 倍，隐藏维度

模型	多样性	MAUVE	连贯性
GPT2-small	$88.13_{\pm 0.12}$	$86.62_{\pm 1.10}$	$53.77_{\pm 0.29}$
TRIME	$88.53_{\pm 0.14}$	$86.76_{\pm 0.58}$	$57.58_{\pm 1.05}$
HISTALIGN	$\mathbf{90.07}_{\pm 0.19}$	$\mathbf{87.46}_{\pm 0.80}$	$\mathbf{61.30}_{\pm 0.15}$
GPT2-large	$88.82_{\pm 0.07}$	$86.18_{\pm 0.94}$	$52.39_{\pm 0.10}$
TRIME	$\mathbf{90.70}_{\pm 0.08}$	$\mathbf{87.27}_{\pm 0.85}$	$53.11_{\pm 0.19}$
HISTALIGN	$89.41_{\pm 0.08}$	$86.83_{\pm 1.02}$	$\mathbf{53.51}_{\pm 0.05}$

表3：开放式生成的自动评估结果。数值为3次运行的平均值 ± 95%置信区间。

流畅度			连贯性		
Win↑	Tie	丢失↓	Win↑	Tie	丢失↓
46.33	36.33	17.33	48.33	32.00	19.66

表 4：开放式生成的人工评估结果。我们对 HISTALIGN 与 TRIME 进行成对比较（“胜出”表示人类更偏好我们的 HISTALIGN 而非 TRIME），并展示被评定为连贯且流畅的段落百分比。HISTALIGN 在流畅度和连贯性方面显著优于 TRIME（$p < 0.05$）。

大 3.2 倍（$1280 \rightarrow 4096$）。因此，正如第 4.1 节所述，由于隐藏维度更大，该模型缓解了 softmax 瓶颈。

然而，我们仍然观察到 LLaMA2 存在这两个问题。首先，softmax 瓶颈问题仍然存在，因为其输出对数概率矩阵 $\mathbf{A}$ 的秩仍受限于其 4096 的隐藏维度；我们发现其经验秩为 3332。这意味着从理论上讲，它的表达能力仍不及高度依赖上下文的自然语言。其次，TRIME 仍然无法充分利用缓存，也就是说，错位问题依然存在。如表2所示，在仅缓存设置下，TRIME 的 Acc@{2,5,10} 准确率为 0%，这表明对于更大的语言模型，错位问题甚至更加明显：由于词元logits 的表现已经足够好，模型不再学习使用缓存。然而，如表中所示，我们的训练目标可以强制模型使用本地缓存并达到 100% 的准确率，这与我们从较小模型中得出的发现一致。

这两个问题的存在表明，语言模型的上下文依赖仍有改进空间，因为在充分利用缓存方面，HISTALIGN 的表现优于 TRIME。

6.2 开放式生成中的连贯性

提示续写的结果见表3。在两种模型规模下，我们都观察到 TRIME 提升了连贯性，而 HISTALIGN 带来的提升更大。HISTALIGN 的效果在较小模型上尤为显著：与原始模型相比，连贯性提高了 7.5 分；与 TRIME 相比，则提高了 3.7 分。这验证了我们的假设，即 HISTALIGN 能够提升语言模型的连贯性。观察 MAUVE 指标时，使用小模型的情况下，HISTALIGN 相较于 GPT2 和 TRIME 分别提高了 0.8 分和 0.7 分。在大模型上，尽管 TRIME 取得了最佳的性能方面，HISTALIGN 仍然比原始模型提高了 0.7 分。多样性也呈现出类似的趋势。总体而言，HISTALIGN 在保持相近多样性和 MAUVE 的同时，提高了连贯性。

除自动评估外，我们还进行了**人工评估**，结果如表 4所示。在流畅度和连贯性方面，人工评分者更偏好 HISTALIGN 生成的续写，而不是 TRIME 生成的续写。这证实了自动评估中的观察结果，即 HISTALIGN 确实尤其改善了连贯性。

6.3 抽象式摘要中的忠实度

摘要结果如表5所示。TRIME 在 XSum 数据集上提高了相较于基线的忠实度，但在 CNN/DM 数据集上的提升并不明显。相比之下，我们提出的 HISTALIGN 方法相较于基线取得了显著提升，尤其是在 DAE 和 P_{ent} 上，这两项指标专门针对幻觉问题。具体而言，在 XSum 数据集上，我们使 FactCC 提高了 0.91 分，DAE 提高了 4.78 分，P_{ent} 提高了 3 分。HISTALIGN 在 CNN/DM 数据集上也改善了各项指标，但提升幅度较小。这表明，使模型能够对输入中的先前上下文给予特定关注，有助于减少幻觉。

我们注意到，HISTAL-IGN 的 ROUGE-L 分数低于原始模型。这种 ROUGE-忠实度权衡已被许多先前的研究观察到（Chen 等人，2021年；Kryscinski 等人，2020年；Wan 和 Bansal，2022年；Wan 等人，2023年），因为参考摘要本身包含幻觉，因此与更忠实的生成摘要并没有很高的重叠度。

为确认这一点，我们开展了**人工评估**。结果见表6。HISTALIGN 的忠实度得分最高，且在统计学上显著优于 BART。这证实了我们从自动评估指标结果中观察到的现象

	XSum				CNN/DM			
模型	ROUGE-L	FactCC	DAE↓	P_{ent}	ROUGE-L	FactCC	DAE↓	P_{ent}
BART	36.41	22.16	67.96	72.72	**30.63**	72.63	6.98	93.53
TRIME	**36.50**	22.94	66.34	74.25	30.60	72.65	7.08	93.39
HISTALIGN	35.45	**23.07**	**63.18**	**75.71**	29.96	**74.93**	**5.73**	**93.80**

表5：抽象式摘要任务上的表现。HISTALIGN 在两个数据集上均始终比两种基线方法更能提升忠实度。

模型	忠实度	信息量
BART	19.33	63.67
TRIME	20.00	66.33
HISTALIGN	26.33*	65.33

表6：XSum 上的人工评估结果。* 表示在统计学上显著优于 BART（$p < 0.05$）。忠实度和信息量的 Krippendorff's α 分别为 0.52 和 0.34。

在表5中。尽管信息量略有下降，但这三种方法之间的差异在统计上并不显著。[5] 这表明，ROUGE-L 等自动化指标的下降并不一定意味着信息量减少。

6.4 数据到文本生成中的忠实度

LogicNLG 上的结果如表7所示。与抽象式摘要类似，HISTALIGN 能够提高 LogicNLG 上的忠实度。在四种忠实度指标中，HISTALIGN 的 NLI-Acc、TAPEX-Acc 和 TAPAS-Acc 最高：在 TAPEX-Acc 上，HISTALIGN 相较于 BART 和 TRIME 分别提高了 0.6 和 0.8 个百分点；在 TAPAS-Acc 上，HISTALIGN 相较于 BART 模型提高了 1.74 个百分点。同时，HISTALIGN 获得了最佳 BLEU分数。

模型	BLEU-(1/2/3)	NA	SA	TX	TS
BART	56.27/37.07/25.63	85.46	**53.45**	63.97	63.74
TRIME	56.12/36.84/25.29	84.55	52.85	63.74	65.11
HISTALIGN	**56.65/37.56/26.25**	**85.67**	53.12	**64.58**	**65.48**

表7：LogicNLG（数据到文本生成）在 BLEU 分数、NLI-Acc（NA）、SP-Acc（SA）、TAPEX-Acc（TA）和 TAPAS-Acc（TS）上的表现。HISTALIGN 在 BLEU 和三项忠实度指标（NA、TX 和 TS）上优于两种基线方法。

7 讨论与结论

在本研究中，我们引入了一种新颖的缓存语言模型训练目标 HISTALIGN，通过为缓存组件添加一种利用顺序信息的对比损失，改进了现有的缓存语言模型目标，从而增强了语言模型的上下文依赖。在一个合成数据集上，我们展示了 HISTALIGN 能够有效地从缓存中检索所需的记忆，并突破 softmax 瓶颈。此外，我们还证明了 HISTALIGN 在提升开放式生成的连贯性，以及提高抽象式摘要和数据到文本生成的忠实度方面的有效性。

随着近期人们不断推动模型规模变得更大，我们想强调几点值得注意的趋势，更强大的模型。首先，如表2所示，仅靠注意力机制无法突破 softmax 瓶颈。其次，虽然增大模型规模可以缓解这一瓶颈，但除非模型规模大到足以真正涵盖人类语言的复杂性，否则这一问题仍会持续存在。缓存语言模型是一种轻量级的替代方案，理论上可以突破 softmax 瓶颈，并在实证上改善上下文依赖。

致谢

我们感谢审稿人和 Haw-Shiuan Chang 帮助提供歧义模板数据。本研究得到 NSF-CAREER奖项1846185、NSF-AI Engage Institute DRL-2112635、DARPA机器常识（MCS）资助项目 N66001-19-2-4031 以及彭博数据科学博士奖学金的支持。本文所述观点仅代表作者本人，并不代表资助机构。

局限性

虽然我们聚焦于局部记忆，以表明当前语言模型仍能从更好的局部上下文依赖中获益，但我们的方法也兼容外部记忆，这有望在未来工作中进一步提升 HISTALIGN 的性能。我们使用 GPT2 和 BART 评估 HISTALIGN，这两种模型最多包含 7.74 亿个参数，规模小于参数量可达数十亿的最新大型语言模型。在歧义模板任务上，我们确实证明了这一问题也存在于采用 LLaMA2 7B 模型的近期大语言模型中，并且我们的方法改善了缓存对齐，但我们

[5] 我们使用自助法检验（Efron 和 Tibshirani，1993年）来确定本文中的统计显著性。

希望未来能够探索将这一方法扩展到大型语言模型，并应用于各种任务。我们相信，我们的方法对更大的模型仍然有帮助。不过，由于较大的模型受 softmax 瓶颈的影响较小（第 4.1 节），它究竟能带来多大帮助，是一个值得未来研究的问题。本研究的另一项局限是，由于引入了额外的超参数（间隔的 λ 和对比损失的权重 α），与 TRIME 相比，将我们的 HISTALIGN 训练目标纳入预训练并不那么直接。该训练目标还假设每个词元具有固定的间隔（因此假定每个词元的差异程度相同），这一点可以通过动态调整间隔来改进。尽管微调成本更低，而且我们展示了在微调中使用 HISTALIGN 能带来有效提升，但如何使用 HISTALIGN 对语言模型进行预训练，也是一个值得探索的未来研究方向。

伦理考量

正如 OpenAI 团队所指出的，GPT-2 无法区分事实与虚构，因此无法支持要求生成文本真实可靠的应用场景。此外，GPT-2 反映了训练数据中固有的偏见，因此部署者必须先针对预期应用场景开展相关偏见研究，之后才能部署。尽管我们的 HISTALIGN 提高了 GPT-2 生成文本的连贯性，但上述观点仍然成立。同样，尽管 HISTALIGN 提高了 BART-large 在抽象摘要和数据到文本生成任务中的生成忠实度，但若不进行进一步检查，此类系统仍不能直接部署并用于对事实准确性敏感的场景。

参考文献

Amal Alabdulkarim、Siyan Li 和 Xiangyu Peng。2021年。自动故事生成：挑战与尝试。载于 第三届叙事理解研讨会论文集，第 72–83 页，线上举行。计算语言学协会。

Shuyang Cao 和 Lu Wang。2021年。CLIFF：通过对比学习提升抽象式摘要的忠实度和事实准确性。载于*2021*年自然语言处理实证方法会议论文集，第6633–6649页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Ziqiang Cao、Furu Wei、Wenjie Li 和 Sujian Li。2018年。《忠于原文：基于事实感知的神经抽象式摘要》。载于第三十届第二届 *AAAI*人工智能会议、第三十届人工智能创新应用会议和第八届*AAAI*人工智能教育进展研讨会，AAAI'18/IAAI'18/EAAI'18。AAAI出版社。

Haw-Shiuan Chang 和 Andrew McCallum。2022年。Softmax 瓶颈使语言模型无法表示多模态词分布。载于第*60*届计算语言学协会年会论文集（第*1*卷：长篇论文），第8048–8073页，爱尔兰都柏林。计算语言学协会。

Haw-Shiuan Chang、Zonghai Yao、Alolika Gon、Hong Yu 和 Andrew McCallum。2023年。重新审视指针网络等架构，以高效改进下一词分布、摘要的事实准确性等方面。载于计算语言学协会论文集：*ACL 2023*（*ACL* 论文集）。

Sihao Chen、Fan Zhang、Kazoo Sone 和 Dan Roth。2021年。通过对比候选生成与选择提升抽象式摘要的忠实度。载于*2021*年计算语言学协会北美分会年会论文集：人类语言技术，第 5935–5941 页，线上。计算语言学协会。

Wenhu Chen、Jianshu Chen、Yu Su、Zhiyu Chen 和 William Yang Wang。2020a。基于开放域表格的逻辑自然语言生成。载于第 *58* 届计算语言学协会年会论文集，第 7929–7942 页，线上。计算语言学协会。

Wenhu Chen、Hongmin Wang、Jianshu Chen、Yunkai Zhang、Hong Wang、Shiyang Li、Xiyou Zhou 和 William Yang Wang。2020b。Tabfact：用于基于表格的事实验证的大规模数据集。载于国际学习表征会议。

Bradley Efron 和 Robert J. Tibshirani。1993年。《自助法导论》。《统计学与应用概率专论》第57卷。Chapman & Hall/CRC，美国佛罗里达州博卡拉顿。

Julian Eisenschlos、Syrine Krichene 和 Thomas Müller。2020年。通过中间预训练理解表格。载于计算语言学协会：*EMNLP 2020* 论文集，第 281–296 页，线上。计算语言学协会。

Angela Fan、Mike Lewis 和 Yann Dauphin。2018年。层次化神经故事生成。载于计算语言学协会第*56*届年会论文集（第*1*卷：长篇论文），第889–898页，澳大利亚墨尔本。计算语言学协会。

Tianyu Gao、Xingcheng Yao 和 Danqi Chen。2021年。SimCSE：句子嵌入的简单对比学习。载于*2021*年自然语言处理实证方法会议论文集，第6894–6910页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Tanya Goyal 和 Greg Durrett。2021年。为摘要中的细粒度事实准确性进行标注与建模。载于计算语言学协会北美分会人类语言技术 *2021*年会议论文集，第 1449–1462 页，线上。计算语言学协会。

Edouard Grave、Moustapha M Cisse 和 Armand Joulin。2017a。用于开放词汇在线语言建模的无界缓存模型。神经信息处理系统进展，第 30 卷。

Edouard Grave、Armand Joulin 和 Nicolas Usunier。2017b。使用连续缓存改进神经语言模型。载于国际学习表征会议。

Edouard Grave、Armand Joulin 和 Nicolas Usunier。2017c。使用连续缓存改进神经语言模型。载于国际学习表征会议。

Caglar Gulcehre、Sungjin Ahn、Ramesh Nallapati、Bowen Zhou 和 Yoshua Bengio。2016年。指向未知词。载于计算语言学协会第 *54* 届年会论文集（第 *1* 卷：长篇论文），第 140–149 页，德国柏林。计算语言学协会。

Karl Moritz Hermann、Tomas Kocisky、Edward Grefen- stette、Lasse Espeholt、Will Kay、Mustafa Suleyman 和 Phil Blunsom。2015年。教机器阅读和理解。载于神经信息处理系统进展，第 28 卷。Curran Associates, Inc.

Ari Holtzman、Jan Buys、Li Du、Maxwell Forbes 和 Yejin Choi。2020年。 神经文本退化的奇特案例。发表于 国际学习表征会议。

Sekitoshi Kanai、Yasuhiro Fujiwara、Yuki Yamanaka 和 Shuichi Adachi。2018年。 Sigsoftmax：softmax 瓶颈的重新分析。发表于 神经信息处理系统进展，第 31 卷。Curran Associates, Inc.

Urvashi Khandelwal、Omer Levy、Dan Jurafsky、Luke Zettlemoyer 和 Mike Lewis。2020年。通过记忆实现泛化：最近邻语言模型。载于国际学习表征会议。

Wojciech Kryscinski、Bryan McCann、CaimingXiong 和 Richard Socher。2020年。评估抽象式文本摘要的事实一致性。收录于 *2020*年自然语言处理实证方法会议论文集（*EMNLP*），第 9332–9346 页，线上。计算语言学协会。

Roland Kuhn 和 Renato De Mori。1990年。一种基于缓存的自然语言模型，用于语音识别。*IEEE* 模式分析与机器智能汇刊，12(6):570–583。

Mike Lewis、Yinhan Liu、Naman Goyal、Marjan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020年。BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练。载于计算语言学协会第*58*届年会论文集，第7871–7880页，线上。计算语言学协会。

Quentin Lhoest、Albert Villanova del Moral、Yacine Jernite、Abhishek Thakur、Patrick von Platen、Suraj Patil、Julien Chaumond、Mariama Drame、Julien Plu、Lewis Tunstall、Joe Davison、Mario Šaško、Gun-jan Chhablani、Bhavitvya Malik、Simon Brandeis、Teven Le Scao、Victor Sanh、Canwen Xu、Nicolas Patry、Angelina McMillan-Major、Philipp Schmid、Sylvain Gugger、Clément Delangue、Théo Matus-sière、Lysandre Debut、Stas Bekman、Pierric Cis-tac、Thibault Goehringer、Victor Mustar、François Lagunas、Alexander Rush 和 Thomas Wolf。2021年。数据集：自然语言处理社区图书馆。载于*2021*年自然语言处理经验方法会议论文集：系统演示，第175–184页，线上和多米尼加共和国蓬塔卡纳。计算语言学协会。

Xiang Lisa Li、Ari Holtzman、Daniel Fried、PercyLiang、Jason Eisner、Tatsunori Hashimoto、Luke Zettlemoyer 和 Mike Lewis。2022a。对比解码：将开放式文本生成视为优化问题。

Xiang Lisa Li、Ari Holtzman、Daniel Fried、PercyLiang、Jason Eisner、Tatsunori Hashimoto、Luke Zettlemoyer 和 Mike Lewis。2022b。对比解码：将开放式文本生成视为优化问题。

Chin-Yew Lin。2004。ROUGE：用于自动评估摘要的软件包。载于《文本摘要：拓展方向》，第74–81页，西班牙巴塞罗那。计算语言学协会。

Ao Liu、Haoyu Dong、Naoaki Okazaki、Shi Han 和 Dongmei Zhang。2022a。PLOG：用于逻辑表格到文本生成的表格到逻辑预训练。载于《*2022*年自然语言处理实证方法会议论文集》，第5531–5546页，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

Qian Liu、Bei Chen、Jiaqi Guo、Morteza Ziyadi、Zeqi Lin、Weizhu Chen 和 Jian-Guang Lou。2022b。TAPEX：通过学习神经SQL执行器进行表格预训练。载于国际学习表征会议。

Yixin Liu、Pengfei Liu、Dragomir Radev 和 Graham Neubig。2022c。BRIO：为抽象式摘要理清顺序。载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），第2890–2903页，爱尔兰都柏林。计算语言学协会。

Ilya Loshchilov 和 Frank Hutter。2019年。解耦权重衰减正则化。收录于国际学习表征会议。

Joshua Maynez、Shashi Narayan、Bernd Bohnet 和 Ryan McDonald。2020年。抽象式摘要中的忠实度与事实准确性。收录于计算语言学协会第*58*届年会论文集，第1906–1919页，线上。计算语言学协会。

Clara Meister、Tiago Pimentel、Gian Wiher 和 Ryan Cotterell。2022年。局部典型采样。计算语言学协会会刊，abs/2202.00666。

Stephen Merity、Caiming Xiong、James Bradbury 和 Richard Socher。2017年。指针哨兵混合模型。收录于国际学习表征会议。

Tomas Mikolov、Ilya Sutskever、Kai Chen、Greg S Corrado 和 Jeff Dean。2013年。词语与短语的分布式表示及其组合性。收录于神经信息处理系统进展，第26卷。Curran Associates, Inc.

Sewon Min、Weijia Shi、Mike Lewis、Xilun Chen、Wen tau Yih、Hannaneh Hajishirzi 和 Luke Zettlemoyer。2022年。非参数掩码语言建模。

Feng Nan、Ramesh Nallapati、Zhiguo Wang、Cicero Nogueira dos Santos、Henghui Zhu、Dejiao Zhang、Kathleen McKeown 和 Bing Xiang。2021年。抽象式文本摘要的实体级事实一致性。载于第*16* 届计算语言学协会欧洲分会会议论文集：主要论文集，第 2727–2733 页，线上。计算语言学协会。

Shashi Narayan、Shay B. Cohen 和 Mirella Lapata。2018年。别给我细节，只要摘要！面向主题的卷积神经网络用于极端摘要。载于*2018* 年自然语言处理实证方法会议论文集，第 1797–1807 页，比利时布鲁塞尔。计算语言学协会。

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei- Jing Zhu。2002年。BLEU：一种机器翻译自动评估方法。载于计算语言学协会第*40*届年会论文集，第311–318页，美国宾夕法尼亚州费城。计算语言学协会。

Krishna Pillutla、Swabha Swayamdipta、Rowan Zellers、John Thickstun、Sean Welleck、Yejin Choi 和 Zaid Harchaoui。2021年。Mauve：使用散度前沿衡量神经文本与人类文本之间的差距。载于神经信息处理系统进展，第34卷，第4816–4828页。Curran Associates, Inc.。

Alec Radford、Jeff Wu、Rewon Child、David Luan、Dario Amodei 和 Ilya Sutskever。2019年。语言模型是无监督的多任务学习者。*OpenAI* 博客。

Hannah Rashkin、Asli Celikyilmaz、Yejin Choi 和 Jianfeng Gao。2020年。PlotMachines：通过动态情节状态跟踪进行大纲条件生成。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第4274–4295页，线上。计算语言学协会。

Nils Reimers 和 Iryna Gurevych。2019年。Sentence-BERT：使用孪生 BERT 网络的句子嵌入。载于*2019*年自然语言处理实证方法会议和第九届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集，第3982–3992页，中国香港。计算语言学协会。

Abigail See、Peter J Liu 和 Christopher D Manning。2017年。切中要点：使用指针生成网络进行摘要。载于计算语言学协会第*55*届年会论文集（第*1*卷：长篇论文），第1073–1083页。

Yixuan Su、Tian Lan、Yan Wang、Dani Yogatama、Ling- peng Kong 和 Nigel Collier。2022年。一种用于神经文本生成的对比框架。载于神经信息处理系统进展。

CraigThomson 和 EhudReiter。2020年。评估数据到文本系统准确率的黄金标准方法。载于第*13*届自然语言生成国际会议论文集，第158–168页，爱尔兰都柏林。计算语言学协会。

Hugo Touvron、Louis Martin、Kevin Stone、Peter Al- bert、Amjad Almahairi、Yasmine Babaei、Nikolay Bashlykov、Soumya Batra、Prajjwal Bhargava、Shruti Bhosale、Dan Bikel、Lukas Blecher、Cristian Canton Ferrer、Moya Chen、Guillem Cucurull、David Esiobu、Jude Fernandes、Jeremy Fu、Wenyin Fu、Brian Fuller、Cynthia Gao、Vedanuj Goswami、Naman Goyal、An- thony Hartshorn、Saghar Hosseini、Rui Hou、Hakan Inan、Marcin Kardas、Viktor Kerkez、Madian Khabsa、IsabelKloumann、ArtemKorenev、PunitSingh Koura、Marie-Anne Lachaux、Thibaut Lavril、Jenya Lee、Di- ana Liskovich、Yinghai Lu、Yuning Mao、Xavier Mar- tinet、Todor Mihaylov、Pushkar Mishra、Igor Moly- bog、Yixin Nie、Andrew Poulton、Jeremy Reizen- stein、Rashi Rungta、Kalyan Saladi、Alan Schelten、

Ruan Silva、Eric Michael Smith、Ranjan Subrama- nian、Xiaoqing Ellen Tan、Binh Tang、Ross Tay- lor、Adina Williams、Jian Xiang Kuan、Puxin Xu、ZhengYan、IliyanZarov、YuchenZhang、Angela Fan、Melanie Kambadur、Sharan Narang、Aurelien Ro- driguez、Robert Stojnic、Sergey Edunov 和 Thomas Scialom。2023年。Llama 2：开放式基础模型和微调聊天模型。

Aaron vanden Oord、Yazhe Li 和 Oriol Vinyals。2019年。使用对比预测编码进行表征学习。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Ł ukasz Kaiser 和 Illia Polosukhin。2017年。Attention is allyou need。收录于神经信息处理系统进展，第 30 卷。Curran Associates, Inc.

Oriol Vinyals、Meire Fortunato 和 Navdeep Jaitly。2015年。指针网络。神经信息处理系统进展，第 28 卷。

David Wan 和 Mohit Bansal。2022年。FactPEGASUS：面向抽象式摘要的事实性意识预训练与微调。收录于计算语言学协会北美分会*2022*年会议论文集：人类语言技术，第 1010–1028 页，西雅图，美国。计算语言学协会。

David Wan、Mengwen Liu、Kathleen McKeown、Markus Dreyer 和 Mohit Bansal。2023年。面向抽象式摘要的忠实度感知解码策略。载于计算语言学协会欧洲分会第*17*届会议论文集，第2864–2880页，克罗地亚杜布罗夫尼克。计算语言学协会。

Rongxiang Weng、Heng Yu、Xiangpeng Wei 和 Wei- hua Luo。2020年。致力于提升神经机器翻译的忠实度。载于*2020*年自然语言处理实证方法会议论文集（*EMNLP*），第2675–2684页，线上。计算语言学协会。

Thomas Wolf、Lysandre Debut、Victor Sanh、J ulien Chaumond、Clement Delangue、Anthony Moi、Pier- ric Cistac、Tim Rault、Remi Louf、Morgan Funtow- icz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander Rush。2020年。Transformers：最先进的自然语言处理。载于*2020*年自然语言处理实证方法会议论文集：系统演示，第 38–45 页，线上。计算语言学协会。

Zhilin Yang、Zihang Dai、Ruslan Salakhutdinov 和 William W. Cohen。2018年。突破 softmax 瓶颈：一种高秩循环神经网络语言模型。发表于国际学习表征会议。

Zhilin Yang、Thang Luong、Russ R Salakhutdinov 和 Quoc VLe。2019年。Mixtape：高效突破 softmax 瓶颈。载于神经信息处理系统进展，第 32 卷。Curran Associates, Inc.

Pengcheng Yin、Graham Neubig、Wen-tau Yih 和 Se- bastian Riedel。2020年。TaBERT：针对文本与表格数据联合理解的预训练。收录于计算语言学协会第*58*届年会论文集，第8413–8426页，在线。计算语言学协会。

Dani Yogatama、Cyprien de Masson d’Autu me 和 LingpengKong。2021年。自适应半参数语言模型。计算语言学协会会刊，9:362–373。

Zexuan Zhong、Tao Lei 和 Danqi Chen。2022年。通过记忆增强训练语言模型。载于*2022*年自然语言处理实证方法会议论文集，第 5657–5673 页，阿拉伯联合酋长国阿布扎比。计算语言学协会。

A 实验设置详情

除非另有说明，我们使用 Huggingface 的 Transformers 图书馆（Wolf 等人，2020年）训练模型。我们采用训练器的默认设置，包括 AdamW 优化器（Loshchilov 和 Hutter，2019年）和线性学习率调度器。我们使用混合精度和 DeepSpeed。我们使用显存为 48GB 的 RTX A6000 GPU，以及显存为 80GB 的 A100 GPU。

在进行超参数调优时，我们尝试的学习率为 {1e-5,3e-5,5e-5}，λ 介于 {0.001,0.0001,0.00001} 之间，并为所有任务设置对比权重 {0.5,1.0}。对于 HISTALIGN，我们使用 λ 0.001，并设置对比权重 $\alpha = 1$，除非另有说明。

A.1 歧义模板

该数据集分别包含 122k、250k 和 122k 个样例，用于训练集、开发集和测试集。测试集与训练集没有重叠的对角词。遵循Chang 和 McCallum（2022年）的方法，我们冻结输出词汇表以防止过拟合，并且仅根据最后一个词元（目标词元）计算损失。我们根据验证损失选择模型。对于基于 GPT2-large 的模型，训练原始模型约需 30 分钟；使用 4 块 RTX A6000 时，TRIME 和 HISTALIGN 约需 1 小时。对于基于 GPT2-s mall 的模型，原始模型、TRIME 和 HISTALIGN 的训练时间分别为 10 分钟、15 分钟和 15 分钟。

Task	数据集	学习率	步数/轮数	预热比例	批次大小	最大输入词元数	最大输出词元数
歧义模板		1e-5	5 集。	0.0	256	-	1024
开放式生成	写作提示	5e-5	3 集。	0.2	512	-	1024
摘要	XSum	5e-5	15,000 步	0.0	128	512	64
	CNN/DM	3e-5	20,000 步	0.0	128	512	128
数据到文本	LogicNLG	5e-5	10 个 eps	0.0	64	500	200

表8：所有任务的超参数。

		上下文	这位 兄弟和这位孙女是我的 最喜欢的事物，而且我尤其喜欢
Full	GPT2-large	宾夕法尼亚州，哈里斯，匹兹堡	兄弟、侄女、孙女
	TRIME	宾夕法尼亚州，费城，匹兹堡	兄弟，侄女，孙女
	HISTALIGN	宾夕法尼亚州，休斯顿，哈里斯	兄弟，孙女，侄女
缓存-only	GPT2-large	以及，我，休斯顿	这位，孙女，我
	TRIME	我，和，休斯顿	the，孙女，最爱的人
	HISTALIGN	宾夕法尼亚州，休斯顿，以及	孙女，兄弟，这位

表9：歧义模板的定性示例。对于完整设置和仅缓存设置，HISTALIGN 均会从上下文中检索出正确的两个词元，并将其列为最高预测结果。

A.2 提示续写

WritingPrompts[6]（Fan 等，2018年）的训练集、开发集和测试集分别包含 273k、16k 和 15k 个样例。我们使用完整的训练集和开发集，并从测试集中抽取 5000 个样例用于最终评估，以节省时间。我们先使用不同的目标函数在训练集上训练模型。我们将文本切分为每块 512 个词元。生成时，我们使用 $p = 0.95$ 和三个随机种子 = {0,1,42} 进行核采样解码，并对分数取平均。训练原始小型模型约需 1.5 小时，TRIME 约需 2 小时，HISTALIGN 约需 3 小时。在 2 块 A100 上训练 GPT-2 large、TRIME 和 HISTALIGN 分别约需 6 小时、7 小时和 11.5 小时。

A.3 摘要

XSum 是一个新闻摘要数据集，由 BBC 文章组成，在训练集、开发集和测试集中分别包含 204k/11k/11k 个样本。CNN/DM 由 Dailymail 和 CNN 的文章组成，该数据集在训练集、开发集和测试集中分别包含 287k/13k/11k 个样本。我们使用官方软件包来计算忠实度指标。[7] 我们通过使用 spaCy 提取实体来计算 P_{ent}，并且仅将 [PERSON、FAC、GPE、ORG、NORP、LOC、EVENT] 视为允许的实体类型。我们使用 Huggingface 的 Dataset 图书馆（Lhoest 等人，2021年）用于加载 XSum（Narayan 等人，2018年）和 CNN/DM（Hermann 等人，2015年）数据集。我们使用 Hugging Face 的 Metrics 库来计算 ROUGE 分数。在 4 张 A6000 显卡上，原始模型、TRIME 和 HISTALIGN 的训练在 XSum 数据集上均耗时约 5 小时，而在 CNN/DM 数据集上，orig、TRIME 和 HISTALIGN 的训练均耗时约 4 小时。

A.4 数据到文本

我们遵循Liu 等人(2022a)对数据集进行预处理，例如为表格添加数值预计算。我们使用对比权重 $\alpha = 0.5$。LogicNLG (Chen 等人 , 2020a) 包含 2.8 万个训练样例、4 千个验证样例和 4 千个测试样例。我们使用原始评估脚本计算忠实度指标，并使用原始数据集提供的 BLEU 计算脚本。[8] 在 2 张 A100 上，训练原始模型和 TRIME 耗时 2 小时，训练 HISTALIGN 耗时约 5 小时。

B 人工评估详情

在两项人工评估中，我们都使用亚马逊机械土耳其进行标注。我们对工作者的要求相同：必须来自美国，已获批的 HIT 数量超过 10,000，且批准率高于 98%。

[6] https://github.com/urvashik/knnmt/blob/master/examples/stories/README.md

[7] FactCC: https://github.com/salesforce/factCC。DAE: https://github.com/tagoyal/factuality-datasets。

[8] https://github.com/wenhuchen/LogicNLG/blob/master/evaluate.py

B.1 开放式生成

我们使用亚马逊机械土耳其进行标注，以判断人们更喜欢 TRIME 还是 HISTALIGN 生成的续写。由于 TRIME 在自动指标上的表现更好，因此我们不纳入原始模型。我们选取两者生成文本的字符数差异小于 200 的示例，以确保长度相近（因为较短的文本通常自然会更连贯）。我们针对随机选取的 100 个示例，每个示例收集 3 条标注，共计 300 条标注。我们统计被判定为连贯和／或流畅的文本段落所占的比例。

我们每个 HIT 支付 0.5 美元，平均耗时约为 2.5 分钟，因此时薪为 $\geq$ 12 美元。标注页面示例如图 3所示。

B.2 摘要

我们采用与 Wan 和 Bansal(2022年) 相同的设置，也使用资格测试来评估所选生成摘要的忠实度。只有标注正确的工作人员才能执行实际任务。

我们选取最重要的句子，并用省略号替换相关性较低的句子，以减轻工作人员的负担。对于每个摘要，我们使用 SentenceTransformer 根据句子嵌入的余弦相似度，从文档中选取最相关的十个句子[9] (ReimersandGurevych，2019年），然后将每个摘要中选出的所有相关句子合并并展示出来。

每项任务由三名不同的工作人员完成，我们取其平均值作为本文档的评分。最终评分是所有文档的平均事实性评分。每项任务平均耗时约 2.5 分钟，每项任务支付 0.5 美元，因此时薪为 $\geq$ \$12 每小时。标注页面示例见图 4。

C 歧义模板的定性结果

我们在表9中展示了两个定性示例。我们发现，原始模型和 TRIME 都难以将两个正确的词排在最前面的两个选项中。仅缓存结果也反映了这一点，其中无关词语（如和、我，*the*获得了较高的概率。事实上，原始模型的缓存相似度与 TRIME 的相似度相近，这再次表明，即使使用缓存进行训练，也无法保证记忆得到良好对齐。尽管如此，HISTALIGN 在完整设置和仅缓存设置下，都将两个目标词列为最前面的两个选项，这表明模型通过我们的对比目标从良好对齐的记忆中获益。

D 示例输出

我们在图5中展示提示续写的示例输出，在图6和图7中展示摘要的示例输出，并在图8中展示数据到文本的示例输出。

[9]我们使用 all-mpnet-base-v2 模型。

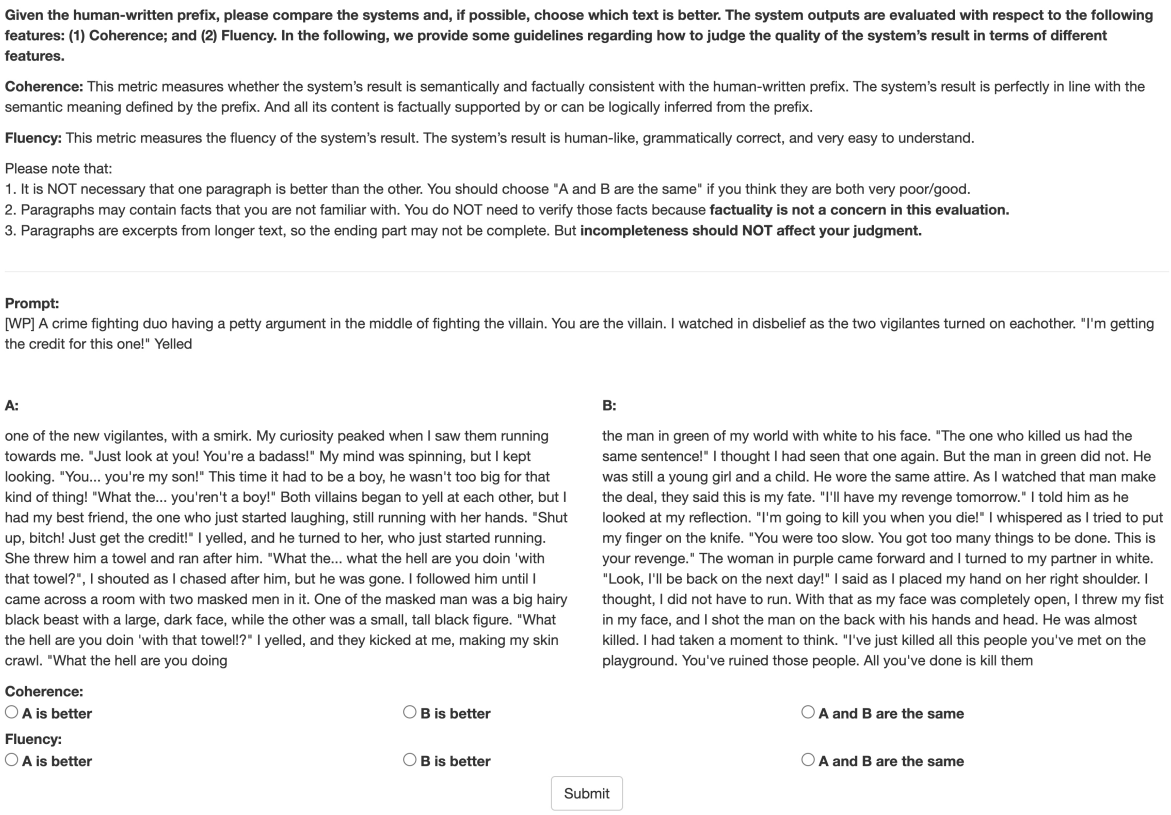
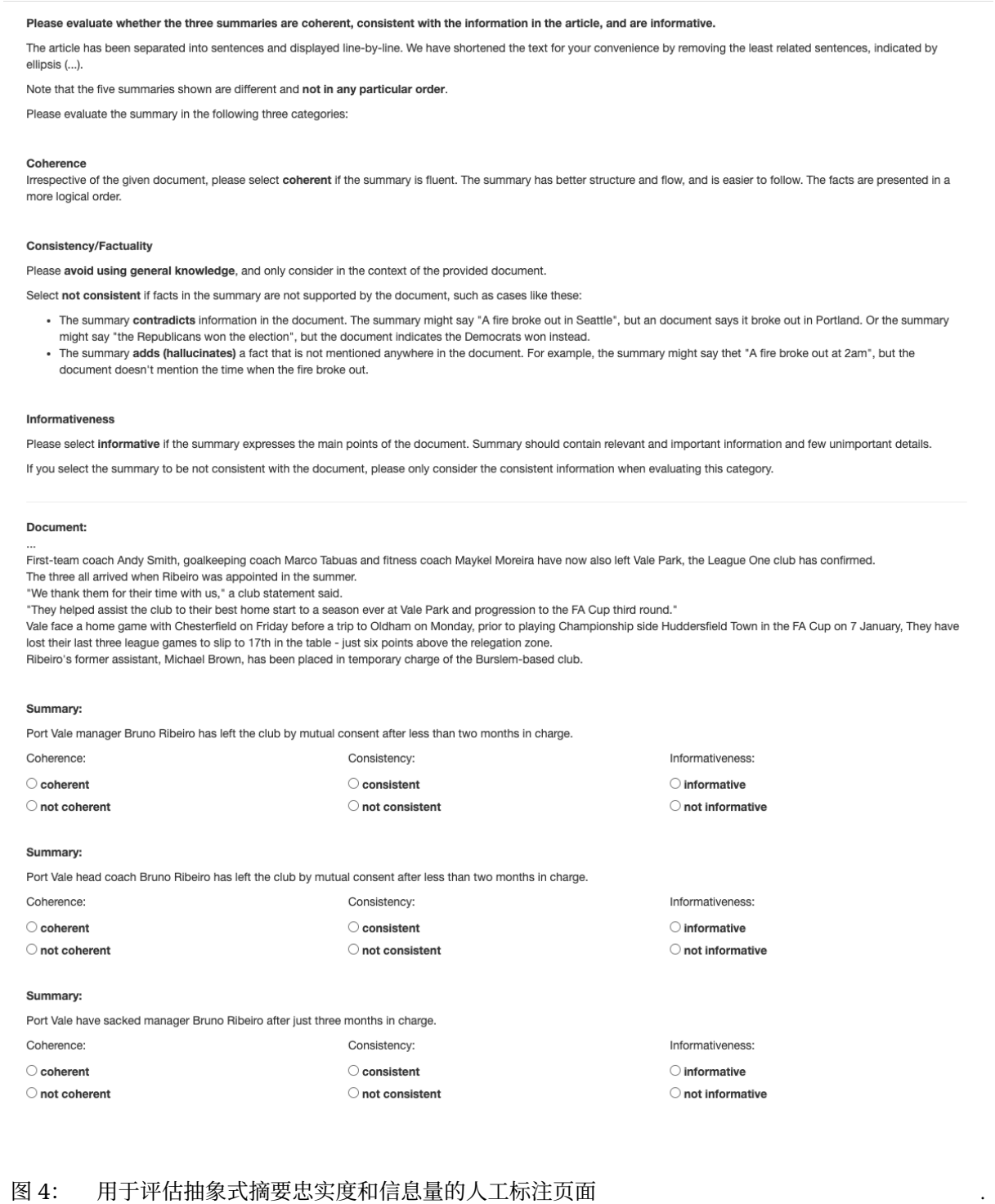


图 3：用于评估提示续写连贯性和流畅度的人工标注页面。



Please evaluate whether the three summaries are coherent, consistent with the information in the article, and are informative.

The article has been separated into sentences and displayed line-by-line. We have shortened the text for your convenience by removing the least related sentences, indicated by ellipsis (...).

Note that the five summaries shown are different and **not in any particular order**.

Please evaluate the summary in the following three categories:

Coherence
Irrespective of the given document, please select **coherent** if the summary is fluent. The summary has better structure and flow, and is easier to follow. The facts are presented in a more logical order.

Consistency/Factuality

Please **avoid using general knowledge**, and only consider in the context of the provided document.

Select **not consistent** if facts in the summary are not supported by the document, such as cases like these:

- The summary **contradicts** information in the document. The summary might say "A fire broke out in Seattle", but an document says it broke out in Portland. Or the summary might say "the Republicans won the election", but the document indicates the Democrats won instead.
- The summary **adds (hallucinates)** a fact that is not mentioned anywhere in the document. For example, the summary might say thet "A fire broke out at 2am", but the document doesn't mention the time when the fire broke out.

Informativeness

Please select **informative** if the summary expresses the main points of the document. Summary should contain relevant and important information and few unimportant details.

If you select the summary to be not consistent with the document, please only consider the consistent information when evaluating this category.

Document:
...
First-team coach Andy Smith, goalkeeping coach Marco Tabuas and fitness coach Maykel Moreira have now also left Vale Park, the League One club has confirmed.
The three all arrived when Ribeiro was appointed in the summer.
"We thank them for their time with us," a club statement said.
"They helped assist the club to their best home start to a season ever at Vale Park and progression to the FA Cup third round."
Vale face a home game with Chesterfield on Friday before a trip to Oldham on Monday, prior to playing Championship side Huddersfield Town in the FA Cup on 7 January, They have lost their last three league games to slip to 17th in the table - just six points above the relegation zone.
Ribeiro's former assistant, Michael Brown, has been placed in temporary charge of the Burslem-based club.

Summary:

Port Vale manager Bruno Ribeiro has left the club by mutual consent after less than two months in charge.

Coherence:	Consistency:	Informativeness:
○ **coherent**	○ **consistent**	○ **informative**
○ **not coherent**	○ **not consistent**	○ **not informative**

Summary:

Port Vale head coach Bruno Ribeiro has left the club by mutual consent after less than two months in charge.

Coherence:	Consistency:	Informativeness:
○ **coherent**	○ **consistent**	○ **informative**
○ **not coherent**	○ **not consistent**	○ **not informative**

Summary:

Port Vale have sacked manager Bruno Ribeiro after just three months in charge.

Coherence:	Consistency:	Informativeness:
○ **coherent**	○ **consistent**	○ **informative**
○ **not coherent**	○ **not consistent**	○ **not informative**

图 4: 用于评估抽象式摘要忠实度和信息量的人工标注页面 .

提示词：每次你死亡时，都会被传送到最终导致你死亡的第一个事件发生的时间。一天，你被一辆汽车撞倒，回到了12年前。这是一段奇异的人生；你不断地设定

GPT2-Large：一路向前，杀死其他人。但这一切已经结束，现在是时候在这个新世界安定下来了。我以前从未真正想过自己的前世，只知道自己回去重新活了一次，但这次却出现了一个问题。我这一生去过很多地方，为什么这次会有这么大的麻烦？这真是太奇怪了，我可以穿越到任何时间、去任何地方，从来不会有这种感觉，但这次却很不对劲。那些汽车和上次一样，就是撞上我的那些车，它们的中间都只是留下了一个洞。我身处一间狭小、废弃的办公室里，墙壁也不完全是用水泥砌成的。我觉得这一点设计得很不错.....

TRIME：.....起来并杀害其他人。但事情已经发生了，现在是时候为这项最新创作画上最后的句号了。坦白说，我真的不知道自己该做些什么。几个月前，我正准备去父母家，路边发生了一起事故。那似乎是某种意外；汽车被一名醉酒驾驶者撞上，随后砸到了我的头上。幸运的是，我还活着。医院告诉我，我只剩下几个月的生命。我倒也能接受这个结果，至少在弄清楚究竟发生了什么之前，我还能继续生活。总之，我现在正等待着与下一次死亡相遇的那一天，而且必须做好应对最坏情况的准备。我可以试着逃避死亡.....

HISTALIGN：忙着四处奔走、阻止各种事情发生，试图避免每一个可能出现的错误。要不是我凡事总是如此谨慎，我或许会是个安分过日子的普通人。像你这样的普通人，不可能回到过去，阻止自己从一开始就死去。任何可能发生的错误都可能当场要了我的命，而这条路走来并不轻松。幸运的是，我恰好出生于1996年7月4日。这些年来，我一直努力做正确的事。我不止一次地死里逃生.....

提示词：鳄鱼 弥诺陶洛斯远征　远征的第一周进展缓慢。地形遍布岩石，崎岖难行，仿佛飞蛾　之母忘了把这片土地整平。弥诺陶洛斯们遇到了最大的麻烦，因为

GPT2-Large：山口的陡坡。当一只牛头怪沿着山口上行时，她会径直穿过山口，仿佛只是路过。那只是众多难以逾越的山脊之一，牛头怪们必须越过它才能穿过山口。它之所以并非完全无法通行，只是因为那些庞大、毛茸茸的野兽占据了山口的一半，迫使许多人离开山区。此外，许多牛头怪穿过山口，也使远征队的减员率不断攀升。尽管如此，牛头怪们还是通过途经.....得知了这处山口

时间：山口的陡坡。当山口到达最高处时，弥诺陶洛斯便无法再通过山口进入下方的山谷。坡实在太陡了。幸运的是，那里有大量岩石，可以用来搭建坡道，这样他们就能在手臂、腿和翅膀完好无损的情况下爬上去。旅程在第三天午餐后开始。小径依然崎岖，许多巨石都将近一英尺高。此时，他们几乎连自己的体重都难以支撑，更不用说装满补给品和设备的行囊了。前方的路上如今布满了小水坑。这是这些.....的典型特征

历史对齐：山口的陡坡。当山口到达最高处时，弥诺陶洛斯便无法再通过山口进入下方的山谷。坡实在太陡了。幸运的是，他们得以扎营，并在最初几天启程。这只是时间问题，他们迟早会意识到自己要去哪里。这个地方远离一切他们熟悉的事物。唯一的光源是他们背上的灯笼，而灯笼只能照亮几分钟。至于他们是如何到达那里的，没人能说得清。他们一直没能找到离开时携带的任何设备.....

图5：开放式提示续写的示例输出。

文档：国际货币基金组织二号人物 David Lipton 周二在华盛顿向全国商业经济学协会发表演讲时，概述了其中一些风险。他说：“国际货币基金组织对全球经济的最新评估再次显示，基线正在走弱。”“我们显然正处于一个微妙的关口。”此前，中国公布的贸易数据弱于预期，出口额较去年同期骤降四分之一。国际货币基金组织已表示，很可能会在 4 月发布下一轮经济预测时，下调目前 3.4% 的全球增长预测。如此黯淡的前景将持续影响押注中国增长故事的企业和行业。阅读 Karishma 的更多文章：散货运输为何值得关注。

BART：国际货币基金组织（IMF）警告称，全球经济正处于一个“关键时刻”。**TRIME**：国际货币基金组织（IMF）警告称，全球经济正处于“危险境地”。**HISTALIGN**：国际货币基金组织（IMF）警告称，全球经济正处于“微妙的关头”。

文档：考文垂大学的斯卡伯勒校区建在小镇原韦珀内斯公园换乘停车场旧址上。目前约有200名学生已在此开始课程，预计最终将有超过2,000名学生在此就读。这座建筑设有工程与科学实验室、模拟法庭和图书馆，是一项耗资5,000万英镑的体育与教育设施的一部分。Craig Gaskell教授表示：“新建的最先进设施正式启用，对我们而言是一个重要的里程碑，也彰显了我们对斯卡伯勒及约克郡海岸地区的承诺。”附近还建成了一所新的大学技术学院，斯卡伯勒竞技足球俱乐部的新建2,000座体育场也正在该旧址上施工。考文垂大学还在伦敦利物浦街站附近设有校区，并于近期宣布将于2017年9月在达格纳姆开设校区。

BART：一所大学在北约克郡正式开设了其首个校区。**TRIME**：一所新大学校区已在北约克郡正式启用。**HISTALIGN**：约克郡海岸的一所大学校区已向公众开放。

图 6：XSum 摘要示例输出。

文档：拜仁慕尼黑将在本赛季最后阶段缺少多面手左后卫大卫·阿拉巴，他因膝关节韧带撕裂预计将缺阵约七周。本赛季早些时候，阿拉巴曾因右膝韧带部分撕裂缺阵三个月。周二，他代表奥地利队出战，在与波斯尼亚队1比1战平的比赛中受伤。“我感到非常失望，”这位22岁的球员在俱乐部官网上表示，“但我有一个远大的目标，那就是赶上赛季收官阶段的比赛。”在奥地利队与波斯尼亚和黑塞哥维那队1比1战平的比赛中，大卫·阿拉巴遭到埃米尔·斯帕希奇的铲球。上周五对阵列支敦士登的比赛中，马克·扬科（左）进球后与马丁·哈尼克和阿拉巴一同庆祝。拜仁表示，阿拉巴预计将在大约一个月后开始康复治疗。自伤愈复出以来，阿拉巴一直为拜仁保持着出色的状态，并凭借一些精彩进球帮助球队稳步朝着三冠王的目标迈进。目前在德国足球甲级联赛中排名榜首、还剩八轮比赛的拜仁，已经晋级德国杯八强，并将在本月晚些时候的欧洲冠军联赛四分之一决赛中迎战波尔图。本赛季，阿拉巴（左）一直为拜仁表现出色，这支巴伐利亚豪门正全力冲击三冠王。

BART：大卫·阿拉巴因膝关节韧带撕裂，将缺阵约七周。周二，阿拉巴代表奥地利出战，在与波斯尼亚队的比赛中以1比1战平，并在比赛中受伤。本月晚些时候，拜仁慕尼黑将在欧洲冠军联赛四分之一决赛中迎战波尔图。

TRIME：大卫·阿拉巴因膝关节韧带撕裂，将缺阵约七周。周二，阿拉巴代表奥地利出战，在与波斯尼亚队的比赛中以1比1战平，并在比赛中受伤。本月晚些时候，拜仁慕尼黑将在欧洲冠军联赛四分之一决赛中迎战波尔图。这位22岁的球员预计将在约一个月后开始康复训练。

HISTALIGN：大卫·阿拉巴因膝关节韧带撕裂，将缺阵七周。这位奥地利后卫在球队与波斯尼亚和黑塞哥维那队1比1战平的比赛中受伤。本月晚些时候，拜仁慕尼黑将在欧洲冠军联赛四分之一决赛中迎战波尔图。

Document：Derry Mathews在利物浦以一致判定胜利击败Tony Luis，夺得世界拳击协会轻量级临时拳王头衔。Mathews此前因疾病两次遇到Richar Abril退出拳击比赛的情况，因此被剥夺拳击腰带；就在本周，替补人选Ismael Barroso也因未能取得签证而被迫退出。周六晚，Derry Mathews在回声竞技场战胜Tony Luis后庆祝。Mathews与家人一同庆祝夺得世界拳击协会轻量级临时拳王头衔。Luis在最后一刻被临时征召上场，这位加拿大人凭借积极进取的风格赢得了许多拳迷的喜爱，但未能获得裁判的认可；Mathews最终以114比112、114比112、115比112的比分获胜。Mathews开局更为出色，以刺拳压制Luis，并在第二回合结束时打出一记漂亮的勾拳。Mathews在利物浦的世界拳击协会轻量级临时拳王头衔拳击比赛中打出一记右拳。Luis（右）是在Richar Abril和Ismael Barroso双双退出后临时顶替上场的。利物浦人Mathews仍在很大程度上掌控着拳击比赛，但Luis也展现出十足的侵略性，不断前压，让Mathews难以应对，并在第六回合结束时打出一记漂亮的勾拳。Luis的压迫和全面进攻开始让Mathews吃不消，后者越来越频繁地退到围绳边；第八回合，两人都被扣了一分，Mathews因击打腰带以下，Luis则因使用头部攻击。最后几个回合中，Luis逐渐占据上风，看起来追回了一些比分，但仍不足以扭转局势。在一场激烈的较量后，Mathews最终以微弱优势获胜。

BART：Derry Mathews 在周六晚于利物浦击败了 Tony Luis。Mathews 眼看着 Richar Abril 两次因疾病退出拳击比赛，因而被剥夺了拳击腰带。Ismael Barroso 因未能获得签证，被迫退出。

TRIME：Derry Mathews 周六晚在利物浦击败了 Tony Luis。Mathews 目睹 Richar Abril 两次因疾病退出拳击比赛。Ismael Barroso 因未能获得签证而被迫退出。Luis 在最后一刻被临时安排参赛，但未能赢得裁判的青睐。

HISTALIGN：Derry Mathews 在回声竞技场通过一致判定胜利击败了 Tony Luis。Mathews 眼看着 Richar Abril 和 Ismael Barroso 因疾病和疾病退出了拳击比赛。裁判判定这位加拿大人以 114-112、114-112 和 115-112 获胜。利物浦人 Luis 在两人退出后于最后一刻被征召参赛。

图7：CNN/DM 摘要的示例输出。

“1928年维多利亚足球联赛赛季”表格：

主队得分	客队	客队得分
11.12 (78)	圣基尔达	21.11 (137)
8.9（57）	吉朗	8.8（56）
11.15 (81)	里士满	10.13（73）
22.17 (149)	山楂	11.13 (79)
18.18（126）	菲茨罗伊	11.13（79）
11.17（83）	埃森登	18.11（119）

参考：圣基尔达在1928年维多利亚足球联赛赛季中作为客队取得了最高比分

BART：霍桑在1928年维多利亚足球联赛赛季中取得了所有球队中最低的客队比分
时间：吉朗的得分最低，为 8.8（56），而霍索恩的得分最高，为 11.13（79）
历史对齐：圣基尔达是 1928年维多利亚足球联赛赛季中得分最高的客队

“1893年英格兰板球赛季”的表格：

name	比赛
John Hearne	29
Tom Richardson	23
Johny Briggs	28
Arthur Mold	28
Bill Lockwood	27

参考：John Hearne 的比赛场次比其他任何球员都多，共参加了 20 场 9

BART：Bill Lockwood 和 Arthur Mold 的比赛场次相同
TRIME：Bill Lockwood 和 Arthur Mold 在 1893年英格兰板球赛季 都参加了 27 场比赛
HISTALIGN：John Hearne 获得了最多的 Match，共 29 次

图 8：数据到文本生成的示例输出。