# GLM: General Language Model Pretraining with Autoregressive Blank Infilling

**摘要**

目前已有各种类型的预训练架构，包括自动编码模型（如 BERT [1]）、自回归模型（如 GPT [2]）和编码器-解码器模型（如 T5 [3]）。然而，在自然语言理解、无条件生成和有条件生成这三大类任务中，没有一个预训练框架能在所有任务中表现最佳。为了应对这一挑战，本文提出了基于自回归空白填充的通用语言模型（GLM）[4]，使得模型兼具有自然语言理解和文本生成的能力。本工作运用 GLM 系列中的 ChatGLM3 进行 AI 生成新闻检测下游任务，在该任务上，使用本工作所提出的方法的性能优于传统的文本分类模型 TextCNN 以及 Bert 模型，并远好于仅仅进行数据微调的 ChatGLM3。

**关键词：** GLM；微调；AI 生成新闻检测

## 1 引言

现有的预训练框架可分为三个系列：以 GPT 为代表的自回归模型、以 BERT 为代表的自编码模型和以 T5 为代表的编码器-解码器模型。自回归模型虽然在长文本生成方面取得了成功，并且在扩展到数十亿参数时显示出了少样本学习能力，但其固有的缺点是单向注意机制，无法完全捕捉自然语言理解任务中上下文单词之间的依赖关系。自动编码模型采用双向注意力机制，获取前后都相互关联的语义信息。其所获得的上下文表征适合自然语言理解任务，但不能直接用于文本生成。编码器-解码器模型结合了编码器以及解码器，其中对编码器采用双向注意，对解码器采用单向注意，两者之间采用交叉注意。而作为编码器解码器模型的代表的 T5 模型虽然统一了自然语言理解和生成任务，但需要更多参数才能与基于 BRET 的模型和 GPT 系列的模型相媲美。它们通常用于条件生成任务，如文本摘要和应答生成。这些预训练框架都不够灵活，无法在所有 NLP 任务中发挥竞争力。在本文中提出了一种基于自回归空白填充的预训练框架 GLM。按照自编码的思路，从输入文本中随机抽取出连续的标记片段，然后按照自回归预训练的思路，训练模型依次重建这些片段（见图 1），并提出二维位置编码以及允许以任意顺序预测片段来改进空白填充预训练。经验表明，在参数数量和计算成本相同的情况下，GLM 在 SuperGLUE 基准测试中以较大优势明显优于 BERT，在类似大小的语料库上进行预训练时，GLM 的表现也优于 RoBERTa [5] 和 BART [6]。在参数和数据较少的自然语言理解和生成任务上，GLM 也明显优于 T5。

本工作运用 GLM 系列中的 ChatGLM3 进行 AI 生成新闻检测下游任务，将所获取的待检测新闻与真实新闻之间所存在的幻觉特征以及情绪特征作为 ChatGLM3 的额外输入信息，

辅助 ChatGLM3 进行 AI 生成新闻检测任务的判断，在该任务上，使用本工作所提出的方法的性能优于传统的文本分类模型 TextCNN [7] 以及 Bert 模型，并远好于仅仅进行数据微调的 ChatGLM3。

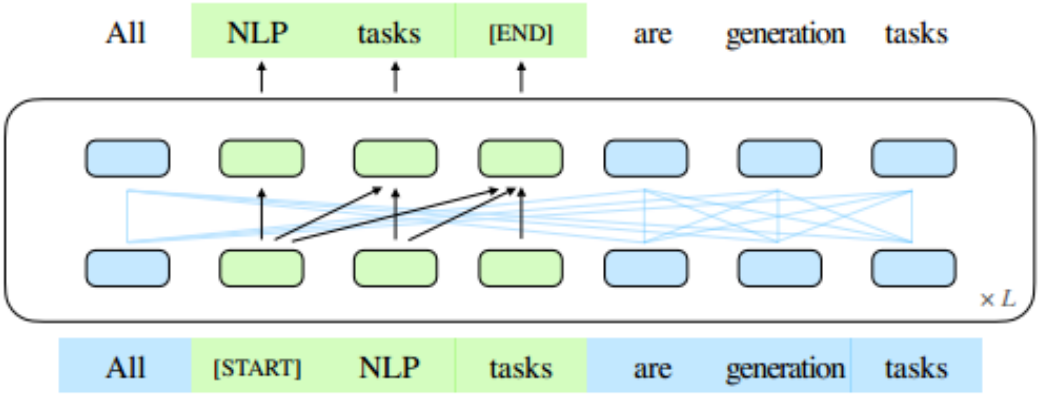


图 1. GLM 重建片段示意图

## 2 相关工作

### 2.1 自然语言处理（NLP）

自然语言处理（NLP）是计算机科学与人工智能领域中的一个重要分支，致力于使计算机能够理解、处理和生成人类语言。NLP 的目标是建立能够处理自然语言文本的智能系统，使其具备与人类语言相似的理解和表达能力。这包括了一系列任务，如自然语言理解、条件生成、无条件生成，具体可细分为文本分类、命名实体识别、情感分析、机器翻译等。随着深度学习等先进技术的发展，NLP 取得了显著的进展，模型如 GPT、BERT 等在各种任务中表现卓越。NLP 的应用涵盖广泛，包括搜索引擎、虚拟助手、智能翻译等领域，为人机交互和信息处理提供了强大的工具。

### 2.2 Transformer 架构

Transformer [8] 架构是一种在自然语言处理领域取得巨大成功的模型。最初由 Vaswani 等人提出，摒弃了传统的循环神经网络（RNN）结构，采用自注意力机制实现对序列数据的建模。Transformer 由编码器和解码器两部分，每个都由多个层叠加而成。自注意力机制使得模型能够在输入序列的不同位置关注不同部分的信息，从而更好地捕捉上下文关系。并且该设计使得它能够并行计算，极大地提高了训练速度。Transformer 在各种任务中表现优异，尤其是在机器翻译、文本生成和语言建模等领域，已成为当今深度学习领域不可或缺的重要组成部分。

### 2.3 预训练语言模型

Transformer 的成功启示了后续预训练语言模型的发展，预训练语言模型是近年来自然语言处理（NLP）领域的关键技术之一，通过在大规模文本语料上进行初始训练，使模型学会

通用的语言表示。这种模型的核心思想是在无监督环境下预先学习丰富的语言知识，然后在特定任务上进行微调，以适应具体的应用场景。其中，BERT 和 GPT 是备受关注的代表性模型，这些模型进一步推动了自然语言处理领域的进步，为构建更智能、适应性更强的自然语言处理系统奠定了基础。

# 3 本文方法

## 3.1 本文方法概述

如图 2所示，GLM 的预训练的设计是首先将一个输入文本序列，其中每一个 x 代表一个 token，从这个序列中随机采样一些片段，将抽取出来的这一些片段随机打乱顺序作为 part B 部分，之所以打乱顺序是不想让顺序造成本身的一个训练上的数据的影响，想完整的捕捉到不同片段之间的依赖关系。然后在原来的文本对应位置用 mask 特殊字符进行替换，作为 part A 部分。将获得的 PartA 和 Part B 拼接在一起作为 GLM 的输入，对于 PartA 部分做像 Bert 一样的双向注意力机制，并且所编码得到的信息相较单项注意力更为丰富，在 partB 部分对于每一个抽取出来的 span 前加 Start 特殊字符做 GPT 式的自回归的预测，即使用单项注意力根据上文信息预测下一个词的形态。结合了自回归以及自编码两种形式，但并不是像 T5 一样的 Encoder-Decoder 架构，对于输入的文本先进 Encoder 再进 Decoder。GLM 中采用的是一种横向的拼接，其中 Encoder 与 Decoder 共享参数。

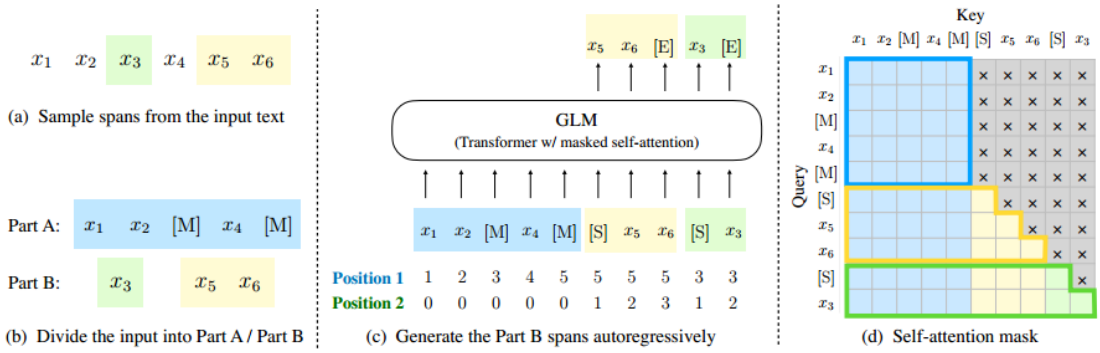


图 2. GLM 预训练结构

## 3.2 多任务预训练

如图 3所示通过改变于 part A 部分所抽取的片段的的长度和数量，假设抽取的 span 长度为 1，上述的模型结构就转换为了完形填空式的 bert 结构，从而适用于自然语言理解任务。假设所输入的文本分为 text1 与 text2 两部分，对于条件生成任务，可以通过抽取 text2 部分，让 GLM 根据 tex1 部分来生成 text2 部分。对于无条件生成任务，即可通过将所有文本内容全部抽取，使得模型需要从无到有的进行生成任务。通过采用如下 (1) 所示的自回归的空白填充目标函数进行预训练即可使模型能够完成”自然语言理解”，”条件生成”，“无条件生成”三类任务的融合。

其中公式 (1) 中 Z 表示所抽取出的所有片段的集合，m 代表所抽取出的片段所包含的 token 的数量，Zm 表示该集合中的长度为 m 的序列。Xcorrupt 代表被 Mask 特殊字符所替

代后的文本序列，即 Part A 部分。S 表示所抽取出的片段序列，该式旨在根据 Part A 部分以及先前生成的文本生成下一个 token 的概率最大化。

$$\max_{\theta} \mathbb{E}_{\boldsymbol{z} \sim Z_m} \left[ \sum_{i=1}^{m} \log p_\theta(\boldsymbol{s}_{z_i} | \boldsymbol{x}_{\text{corrupt}}, \boldsymbol{s}_{\boldsymbol{z}_{<i}}) \right] \quad (1)$$

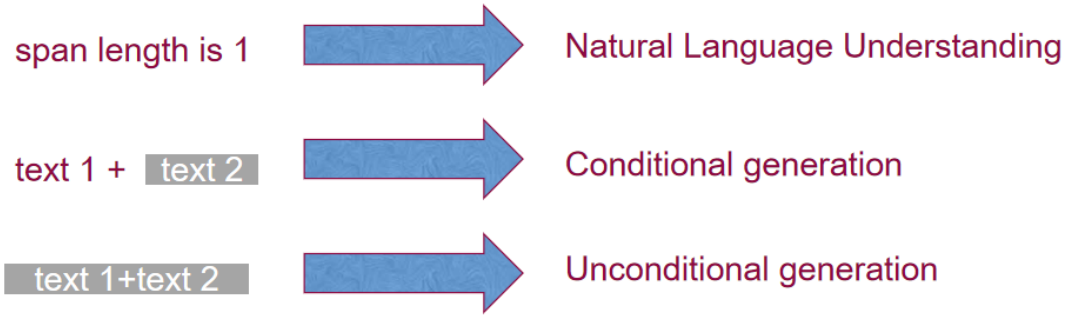


图 3. 多任务预训练

### 3.3 微调

对于下游自然语言理解任务，线性分类器将预训练模型生成的序列或标记的表示作为输入，并预测正确的标签。这种做法与生成式预训练任务不同，导致预训练和微调之间不一致。通过使用空白填充的生成任务重新设计了自然语言理解中的分类任务，如图 4任务的设计，将分类任务范式转换为生成式任务，实现分类任务、生成任务的统一。

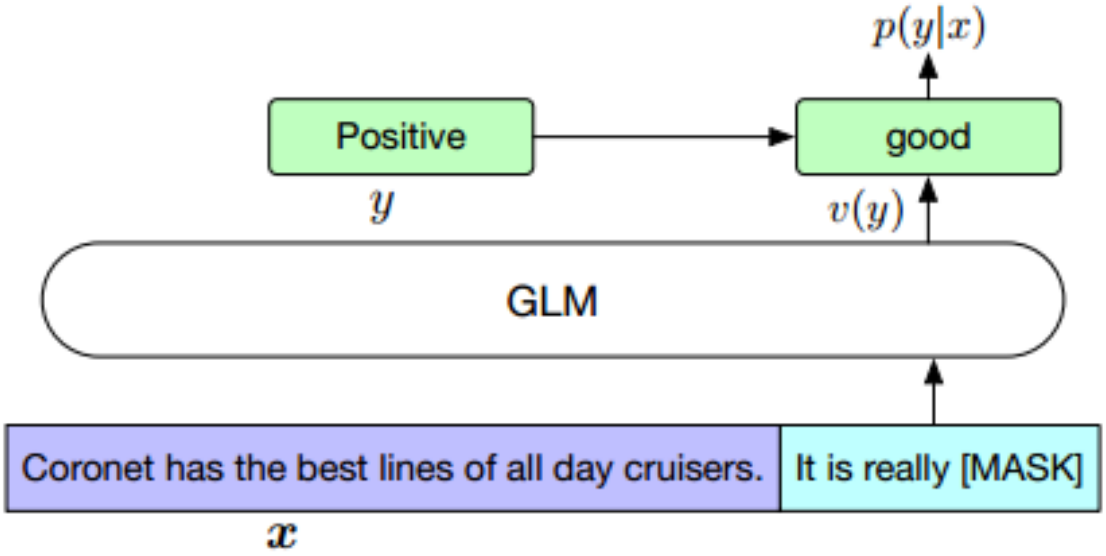


图 4. 转变分类任务范式的微调

### 3.4 二维位置编码

如图 2(c) 部分所示，本文提出了二维位置编码来对位置信息进行编码。具体来说，每个标记都有两个位置 id。第一个位置 id 代表损坏文本 Xcorrupt 中的位置。对于被抽取的片段，它是相应 MASK 字符的位置。第二个位置 id 代表片段内位置。对于 part A 的标记，其第二个位置 id 为 0；对于 part B 的标记，其范围从 1 到片段的长度。这两个位置 id 通过可学习的嵌入表被投射到两个向量中，这两个向量都会被添加到输入的标记嵌入中。这种设计比较适合下游任务，因为这种编码方式确保了模型不知道所抽取掉的片段的长度，而下游任务通常事先不知道生成文本的长度。

## 4 AI 生成新闻检测

### 4.1 实验背景

在数字时代，随着互联网的快速发展，互联网是大众获取新闻内容的主要渠道。然而，网络上的各种信息良莠不齐、鱼龙混杂，网络谣言与虚假内容严重影响了网络文明环境。大模型作为强大的文本生成工具，可能被恶意攻击者用于生成虚假新闻，造成社会恐慌。因此，亟需从海量新闻中高效准确地识别出 AI 生成新闻，从而帮助控制 AI 生成新闻的传播。在本实验中使用 GLM 系列中的 ChatGLM3 进行 AI 生成新闻检测下游任务。

### 4.2 实验调研

大语言模型生成的新闻与新闻工作者撰写的新闻相比有两个显著特点，一个是更可能存在幻觉，可能产生偏离事实或包含编造信息的内容，二是不如人类撰写的文章情感丰富，更正式。写作更有条理且中立，提供的偏见和有害信息较少。而目前相关领域的研究主要集中在多模态 AI 虚假新闻检测中，主要是利用视觉特征提取器提取视觉信息，利用文本特征提取器提取文本特征，之后将视觉信息和文本信息进行匹配，进行虚假新闻检测。

### 4.3 实验环境搭建

torch==2.0.1+cu11.3，transformers==4.30.2 显卡采用 4090。

### 4.4 实验数据集

以往的关于新闻的数据集大多都是做新闻摘要方面的任务，因此在新闻摘要任务数据集中选取了 CNN/DailyMail 数据集，如图 5所示，该数据集分为新闻以及摘要两部分，通过 Chatglm3 使用如下 Prompt 对摘要进行扩写，共生成了 5000 条 AI 生成新闻，如图 6所示。再从 CNN/DailyMail 中随机选取 5000 条新闻，总计 10000 条共同构成本次实验的数据集。

数据生成Prompt:

You are a professional news writer who specializes in expanding news summaries into English-language news articles that maintain the engaging and informative style of CNN reporting while meeting the needs of a diverse audience. Your task is to write an 800-word English-language news article based on the summary provided below: #highlight

| article<br>string | highlights<br>string |
| --- | --- |
| LONDON, England (Reuters) -- Harry Potter star Daniel Radcliffe gains access to a reported £20 million ($41.1 million) fortune as he turns 18 on Monday, but he insists the money won't cast a spell on him. Daniel Radcliffe as Harry Potter in "Harry Potter and the Order of the Phoenix" To the disappointment of gossip columnists around the world, the young actor says he has no plans to fritter his cash away on fast cars, drink and celebrity parties. "I don't plan to be one of those people who, as soon as they turn 18, suddenly buy themselves a massive sports car collection or something similar," he told an Australian interviewer earlier this month. "I don't think I'll be | Harry Potter star Daniel Radcliffe gets £20M fortune as he turns 18 Monday . Young actor says he has no plans to fritter his cash away . Radcliffe's earnings from first five Potter films have been held in trust fund . |

图 5. CNN/DailyMail 数据集

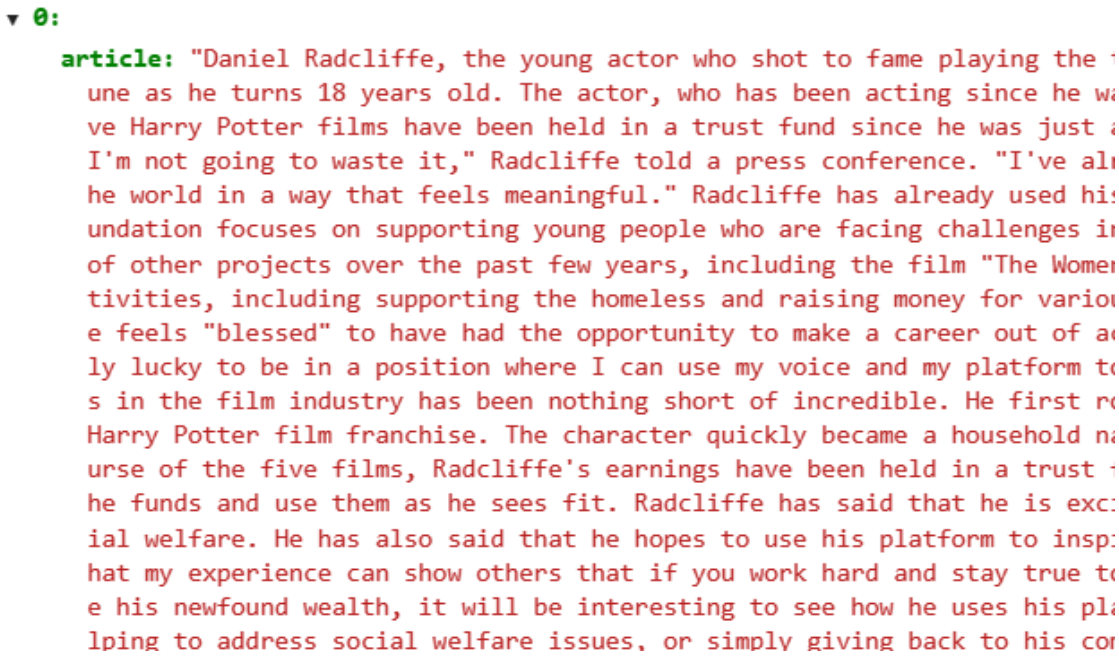


图 6. AI 生成数据集

### 4.5 微调

将数据集以 4：1 的比例构建训练集以及测试集，使用训练集对 chatglm3 进行 P-Tuning 微调迭代 1000 次。数据集中真实新闻平均单词量为 781，AI 生成新闻平均单词量为 714。微

调结果图如图 7所示.

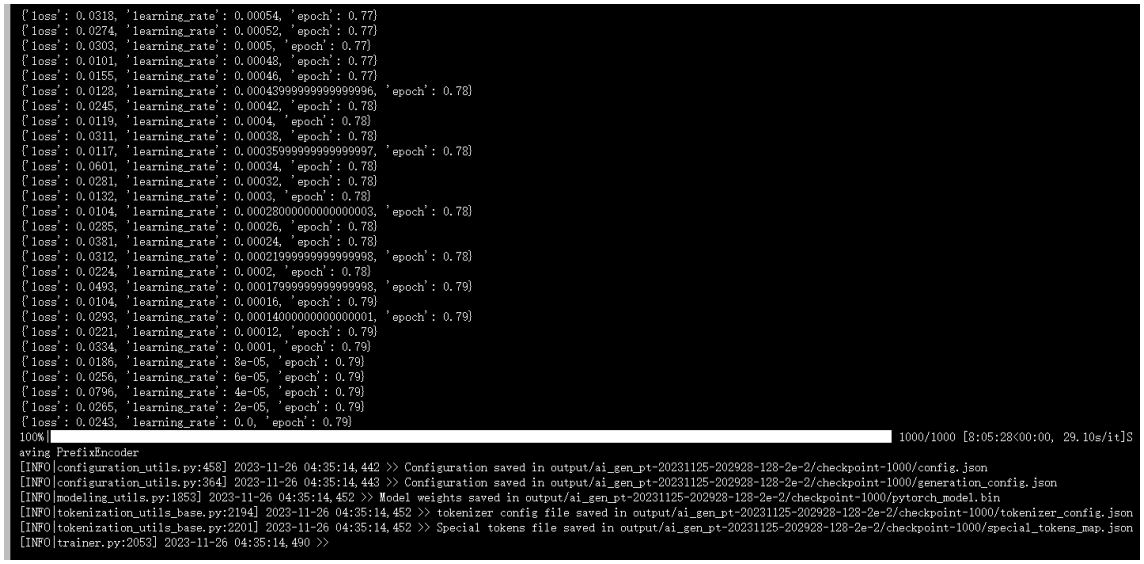


图 7. 微调结果图

### 4.6 创新点

基于大语言模型领域中检索增强技术 (RAG),RAG 技术的核心思想是在生成文本时，先通过检索阶段获取相关信息，然后利用生成模型生成更加准确和丰富的输出。因此将大语言模型与真实新闻之间存在的幻觉特征以及情绪特征作为额外输入信息辅助大语言模型进行 AI 生成新闻检测。

### 4.7 实验

首先基于大语言模型生成的新闻与新闻工作者撰写的新闻相比的更可能存在幻觉的特征，进行基于句子相似性的幻觉检测，即通过计算语言模型为两个句子分别生成的嵌入向量的余弦相似性。如果余弦相似度得分大于某个阈值，则这两个句子被认为是相似的，检索结果中的句子会被预测为有依据的。在本实验中选择 ChatGLM3 的输出向量作为向量模型，考虑到该向量模型存在输出的向量分布不均匀，高频词和低频词处于空间中的不同区域，并且低频词分布稀疏即训练不充分，导致低频词与高频词之间的相似性不适用，因此使用 Bert-whitening 向量技术对该向量模型进行优化。

#### 4.7.1 Bert-whitening

BERT-whitening 方法最早提出于对原 BERT 向量进行空间分布转换，从而提高 BERT 语义向量相似度计算方面的效果，并且降低 BERT 语义向量的维度，并且提高向量检索的效率和更小的内存占用。将该方法应用于对 ChatGLM3 向量进行变换，首先进行去除偏移，即将不同输入序列的向量表示归一化为相同的长度和形式。获得中心化后的向量表示后进行奇异值分解来获得一个正交矩阵，然后将该正交矩阵应用于中心化后的向量表示，从而得到白化后的向量表示。最终使用原始向量表示的协方差矩阵来对白化后的向量表示进行逆变换，从而得到最终的文本向量表示。最终将将数据的协方差矩阵变为对角矩阵，减少特征之间的相关性，提高数据的可解释性和泛化能力，并保证文本表示的质量，具体代码如图 8所示。

```python
def compute_kernel_bias(vecs, n_components=256):

    mu = vecs.mean(axis=0, keepdims=True)
    cov = np.cov(vecs.T)
    u, s, vh = np.linalg.svd(cov)
    W = np.dot(u, np.diag(1 / np.sqrt(s)))
    return W[:, :n_components], -mu

kernel, bias = compute_kernel_bias(vecs)


def transform_and_normalize(vecs, kernel=None, bias=None):

    if not (kernel is None or bias is None):
        vecs = (vecs + bias).dot(kernel)
    return vecs / (vecs**2).sum(axis=1, keepdims=True)**0.5
```

图 8. Bert-whitening 代码

#### 4.7.2 幻觉特征生成

根据同一则新闻事件应当有多份报道的观点，使用 ChatGLM3 对待检测新闻使用如下关键词提取 Prompt 提取关键词，使用关键词对 CNN 网站上进行爬取相关新闻，所爬取的关键词如图 9。由于具有相同关键词的新闻不一定与待检测新闻阐述同一事件，并且 ChatGLM3 本身的向量并不适用于长文档级别，因此使用 ChatGLM3 通过如下总结 Prompt 将相关新闻以及待检测新闻总结为较短的句子，所总结的句子如图 10所示，并将句子间进行余弦相似性比较，选取阈值大于 0.7 的对应新闻作为与待检测新闻描述同一事件的参考新闻。将参考新闻分句存入向量数据库中，使用待检测新闻对该向量数据库进行检索，判断待检测新闻中的描写是否在参考新闻中存在依据，即判断有多少句话在别的新闻报道中也有提及，最终得到关于新闻幻觉的特征信息，如图 11所示。具体流程图如下图 12所示。

关键词Prompt:

Take three keywords from the following news story.The output should consist of only three keywords, not any other extraneous information.#news:

crawl_news0:Daniel Radcliffe has moved on from playing Harry Potter, the beloved character that he previously
's bestselling book series.So when he was asked if he has any plans to be a part of the upcoming Max TV reboot
lished Monday that he "certainly" hasn't had any discussions about it. "I think it's very much like, they're go
so I think it will be very weird for me to show up," Radcliffe said of the idea of him appearing in the series
or continued, "I'm very excited to see what other people do with it," and compared the on-screen retelling of
tically reinterpreted in various ways. "I think the 'Potter' series of books was always going to be bigger than
torch get passed on," he said.Warner Bros. and Max - who are part of Warner Bros. Discovery, like CNN - announ
" television series, with the promise that the new show will be a "faithful adaptation" of Rowling's book seri
2007, is also serving as an executive producer.The series is expected to run for an unprecedented 10 years wit
books and bring Harry Potter and these incredible adventures to new audiences around the world, while the orig
ision series will feature a cast of yet-to-be-announced fresh new faces to take on the roles made famous in th
tarred as Potter, Hermione Granger and Ron Weasley, respectively, between 2001 and 2011.Radcliffe, meanwhile,
e Geraldine Viswanathan, Karan Soni, Jon Bass and Steve Buscemi. Season 4 of the show premieres on July 10.

图 9. 所爬取的新闻示例

总结Prompt:

Summarise the content of the news after # in one sentence.The output should consist of only summary, not any other extraneous information.#news

```
Summary: The young wizard is set to gain access to a reported £20 million fortune as he turns 18 on Monday,
```

图 10. 所生成的总结

```
There is no basis for the presence of 15.384615384615385% in the news to be detected
```

图 11. 所生成的幻觉特征

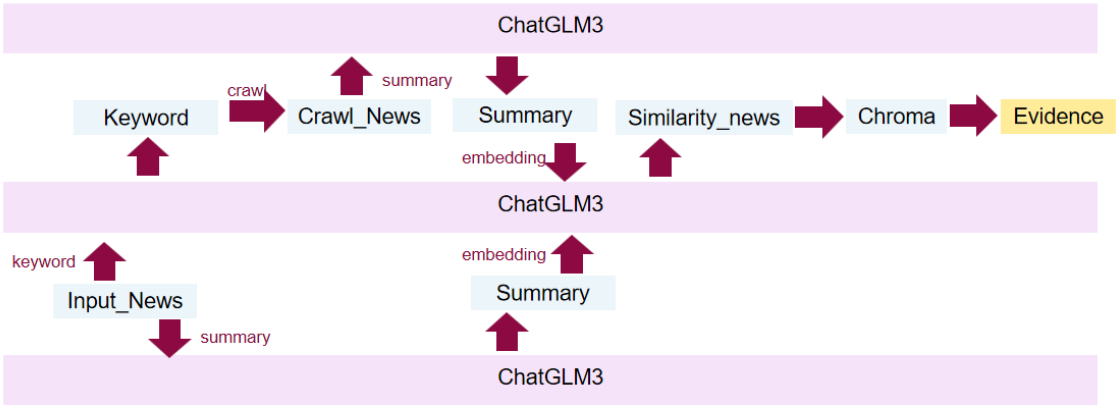


图 12. 流程图

### 4.7.3 情绪特征生成

对于情绪生成部分，通过 ChatGLM3 使用如下 prompt 将待检测新闻与相关新闻进行情绪丰富度的比较，得到情绪特征如图 13所示。

Prompt for Emotion:

You are a psychologist who specialises in analysing the emotions embedded in the news and your task is to conduct an emotional analysis based on the two news stories provided below, comparing and contrasting which one is more emotional:

news1:#

news2:#

```
emotion:The first news story is more emotional because it involves the emotional reaction of Daniel Radcliffe, w
in a new TV series about the same character. The second news story is more factual, as it reports on Daniel Radc
```

图 13. 所生成的情绪特征

#### 4.7.4 注入 Prompt

最终将所获得的幻觉特征以及情绪特征共同注入到如下 Promt 中，将该 Prompt 作为最终输入进行 AI 生成新闻检测，所得结果示例如下图 14所示。

最终Prompt:

Tasks Description：As a text classifier, you need to classify the input text into 'human written' or 'AI generated' based on the following similarity and sentiment analysis from Information.It is known that if the similarity is greater than eighty percent,the news2 is more emotional or the less no basis sentences in the news , it is likely to be written by a human with an additional thirty percent chance.And outputs must be It is ai generated or It is human-written.

information:

#similarity:

#emotion:

news:#news

Based on the provided information, the text can be classified as ' ai generated'

图 14. 所生成的情绪特征

## 5 实验结果分析

将使用该方法构建 Prompt 的 ChatGLM3 与仅仅只是进行微调的 ChatGLM3、TextCNN 以及 BERT 进行对比实验，实验结果如下表 1 所示。可以看到使用该方法构建 Prompt 后，在 AI 生成新闻检测任务上，比仅仅进行微调的 ChatGLM3 有大幅度的提升，而相比传统的文本分类模型 BERT-base-uncased 以及 Textcnn 模型在各指标上也有小幅度的提升，除了在精确度指标上略逊于 BERT-base-uncased 模型。

| Model | Acc | Precision | Recall | F1-score |
|---|---|---|---|---|
| BERT-base-uncased | 0.67 | 0.82 | 0.67 | 0.74 |
| TextCNN | 0.72 | 0.54 | 0.73 | 0.62 |
| ChatGLM3 | 0.59 | 0.38 | 0.65 | 0.48 |
| Ours | 0.75 | 0.69 | 0.82 | 0.76 |

表 1. Caption

## 6 总结与展望

通过研读相关论文和文档，对于 GLM 的架构有了深入的理解，包括模型架构的设计和以及相应 Tokenizer 的实现原理。了解了数据处理和大模型微调过程的基本步骤，为将来在更广泛任务中应用模型奠定了基础。目前对于大语言模型的的理论知识和实践经验仍然较为薄弱，需要进一步学习大语言模型中涌现的新模型以及基本概念和常用技术。由于资源和时间

有限，仅仅在本实验所使用的数据集上进行了训练和验证，对于模型的泛化能力的表现尚未全面评估。未来可能会去考察该模型对于其他新闻数据集的泛化能力，以及情绪特征上的生成。总体而言，这次学习为我打开了大语言模型学习的大门，但也揭示了在知识体系和实践经验上的不足。通过不断学习和实践，将能够更好地应对该领域的挑战，提升自己的技能水平。

## 参考文献


[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

[3] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[4] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[6] Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. 2010.

[7] Bao Guo, Chunxia Zhang, Junmin Liu, and Xiaoyi Ma. Improving text classification with weighted word embeddings via a multi-channel textcnn model. *Neurocomputing*, 363:366–374, 2019.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.