

C 数据集和评估细节

C.1 多任务指令预培训 (MIP)

按照 (Raffel 等人, 2020; Wei 等人, 2022a; Sanh 等人, 2022; Aribandi 等人, 2022) 的做法, 我们在 GLM-130B 的 MIP 训练中包含了一些提示指令数据集, 占训练标记的 5%。T0 数据集的所有提示均来自 PromptSource (巴赫等人, 2022 年), 而 DeepStruct 数据集的提示则是新创建的。它们的组成如表 12 所示, 其中自然语言理解和生成数据集来自 T0 (Sanh 等人, 2022 年) 和 promptsource (Bach 等人, 2022 年), 信息提取数据集来自 DeepStruct (Wang 等人, 2022a)。在 GLM-130B 的训练中, 我们计算出每个数据集中约有 36% 的样本被看过。

T0 最初将数据集分为 1) 多任务提示训练和 2) 零射击任务转移两个部分。我们最初计划只包含 T0 的多任务提示训练部分和 DeepStruct 的训练集 (Wang 等人, 2022a), 但由于失误, 我们在 MIP 中同时包含了多任务提示训练和零射击任务转移部分的数据集, 而排除了 DeepStruct 数据集。这个错误在大约 23k 步时被修正, 我们的模型继续在正确的版本上进行训练。

自然语言理解与生成。我们采用 promptsource (巴赫等人, 2022 年) 中的数据集和相应提示。对于每个数据集中的所有提示样本, 我们设置了每个数据集最多 10,000 个样本的截断值, 并将它们合并为 MIP 数据集。有关提示样本和数据集的详细信息, 请参阅 promptsource 的 GitHub 存储库⁸。

信息提取。DeepStruct 是一种针对信息提取任务的多任务语言模型预训练方法 (Wang 等人, 2022a), 基于该方法的数据集, 我们为其部分数据集创建了指令和提示 (如表 12 所示)。我们将信息提取任务重新制定为指令调整格式, 以便零点泛化到新的提取模式。由于信息提取数据集的数量少于普通语言理解和生成数据集, 因此我们为每个数据集中的所有提示样本设置了最多 20,000 个样本的截断值。对于 KERM 数据集 (Agarwal 等人, 2021 年) 和 PropBank 数据集 (Kingsbury & Palmer), 由于它们的原始大小巨大, 我们从它们的提示样本中各抽取了 50,000 个样本。

C.2 深度结构 MIP 中的数据和提示

DeepStruct (Wang et al., 2022a) 中所有数据集的提示和说明均由作者手动新建。下文附有每个数据集的简介、任务描述和完整提示。为便于模板填充, 所有提示都写入了 Jinja⁹ 模板。当数据集样本以我们的格式提供时, Jinja 引擎会将其渲染为带有说明的提示样本。

对 GLM-130B 的信息提取能力进行更系统的评估将留待今后的工作中进行, 因为这项工作

的重点是 LLM 的训练和设计细节。

C.2.1 对话状态跟踪

我们采用 Multiwoz 2.1 (Eric 等人, 2020 年) 对话状态跟踪数据集。该数据集被重新组合为两个任务, 每个任务都有一个相应的提示:

- **对话状态跟踪:** 要求模型从给定特定时段 (如出租车到达时间和目的地) 列表的对话中提取信息。
- **填空:** 哪个模型应填入一个提供的空格, 并确定没有答案的情况。

⁸<https://github.com/bigscience-workshop/promptsources>
⁹<https://github.com/pallets/jinja>

(对话状态跟踪, 提示 0)

阅读"[用户]"和"[代理]"之间的对话、

{{文本}}

识别并提取与以下类别相关的信息 (从上到下) :

- {{allowed_relations | join("/n- ")}}

([User] ; Y ; Z)" 的形式: || {{format_triple(relations, allowed_relations) | join(" ")}}

(填补空白, 提示 0)

下面是一段对话:

{{文本}}

请回答问题: "[用户]"是否提到"{{allowed_relations[relation_idx].split(': ') | join("'s ")}}" ? 如果有, 请写下对话中的答案; 如果没有, 请回答 "未给出"。

回答: ||| {% if filter_relation(relations, allowed_relations[relation_idx]). len () > 0 %}{{filter_relation(relations, allowed_relations[relation_idx])[0]['tail']]}}{% else %}not given{% endif %}

C.2.2 事件提取

我们采用 ACE05 ([Walker & Consortium, 2005 年](#)) 事件提取数据集, 并沿用了 ([Wadden 等人, 2019 年](#)) 中的设置。该数据集分为两个任务, 分别有以下三个提示:

- **事件参数提取:** 给定文本中的触发器及其参数角色列表, 要求模型从提供的文本中提取参数。
- **参数识别:** 给定一个触发器和某个参数角色, 如果该参数存在于所提供的文本中, 则要求模型提取该参数; 否则, 模型不应生成任何内容。

(事件参数提取, 提示 0)

对于 "事件提取" 任务, 给定一个触发器后, 应根据潜在角色列表提取其相关参数。

给定以下角色列表:

```
- {{shuffle(allowed_arguments[trigger['event_type']].values()) |  
join("\n- ")}}
```

提取触发器"{{trigger['text']}} ({{
allowed_triggers[trigger['event_type']]}})" 的相关参数:

{{文本}}

提取: || {{format_triple(relations, "") | join(" ")}}

(事件论据提取, 提示 1)

测试

1. (事件提取) `{{text}}`

请写下与触发器"`{{trigger['text']}}`

`{{allowed_triggers[trigger['event_type']]}}`" 相关的所有事件参数, 并用"`[]`"标记。

`]`", 给出了以下类别:

```
- {{shuffle(allowed_arguments[trigger['event_type']].values()) |
join("\ n- ")}}
```

请回答: `|| {{format_triple(relations, "") | join(" ")}}`

(论据识别, 提示 0)

让我们提取与事件相关的论据!

在下面的段落中, 类型为"`{{query_arg}}`"的参数与事件触发器"`{{trigger['text']}}`

`{{allowed_triggers[trigger['event_type']]}}`" 相关:

`{{文本}}`

参数应为 (如果找到, 则从上下文中复制; 如果没有, 则不生成): `||`

```
{{filter_type(relations, query_arg) | join(" ")}}
```

C.2.3 联合实体和关系提取

联合实体和关系抽取旨在识别文本中的命名实体, 并判断它们之间的关系。它与知识获取密切相关, 后者的最终目标是将非结构化的网络内容结构化为知识三元组 (如 (伦敦、英国首都))。这项任务可以通过管道框架 (命名实体识别与关系提取的结合) 或端到端训练来完成。

在这项工作中, 我们采用了三个经典的联合实体和关系抽取数据集: CoNLL04 (Roth & Yih, 2004)、NYT (Riedel et al., 2010) 和 ACE2005 (Walker & Consortium, 2005)。在 GLM-130B 中, 我们遵循 (Wang 等人, 2022a) 的观点, 将此类挑战表述为序列到序列的生成, 其中输入为原始文本, 输出为三元组。在此, 我们只对这些数据集执行与关系相关的任务, 而将与实体相关的任务留给命名实体识别部分。

- **关系提取:** 在此, 我们根据候选关系列表, 提取由 "头部实体"、"关系" 和 "尾部实体" 组

成的知识三元组。例如，输入 "在昆明，800 多名师生成立了国立西南联合大学"，模型输出可能是 (国立西南联合大学、成立地点、昆明)。

- **条件关系提取**：给定一个候选关系，判断输入文本是否包含该关系。如果是，则提取所有相关的三元组；如果不是，则不生成。
- **知识槽填充**：从文本中指定某个实体，并要求模型提取所有以该实体为首的三元组。
- **关系分类**：给定文本中的两个实体，要求模型根据候选关系列表判断它们之间的关系。

(关系提取, 提示 0)

您能从句子中找出与"`{{shuffle(allowed_relations) | join(' ', ' ')}}`" 有关的所有三元组吗? 请在

(`X ; Y ; Z`)" 的形状:

```
{{text}} => ||| {{format_triple(relations, allowed_relations) | join(" ")}}
```

(条件关系提取, 提示 0)

以关系"`{{allowed_relations[relations_idx]}}`"为条件, 可以从中提取哪些知识三元组:

`{{文本}}`

请在此写下它们: `{{format_triple(relations, allowed_relations) | join(" ")}}`

(填充知识槽, 提示 0)

`{% if entity_types.在句子中`

`{{文本}}`

`X = "{{entities[entity_idx]}}"` 是"`{{entity_types[entity_idx]}}`"类型的实体。
提取包含"`{{entities[entity_idx]}}`"的所有可能的三元组, 其形式为 (`X ; Y ; Z`)
, 并给出以下候选属性 `Y`:

```
{% for r in allowed_relations %}- {{r}}  
{% endfor %}
```

回答: `||| {% for r in relations %}{% if r['head'][0] == entities[entity_idx] %}{{format_triple([r], allowed_relations) | join(" ")}}{% endif %}{% endfor %}`

(关系分类, 提示 0)

问答

1. 给定候选关系:

```
- {{shuffle(allowed_relations) | join("/n- ")}}
```

下面句子中的"`{{relations[triple_idx]['head'][0]}}`"和"`{{relations[triple_idx]['tail'][0]}}`"是什么关系?

`{{文本}}`

回答: `||| {{relations[triple_idx]['relation']}}`

然而, 现有的实体和关系联合提取数据集的关系模式非常有限。例如, CoNLL04 只包含五种不同的关系; 而最多样化的《纽约时报》数据集则包含 24 个 Freebase 谓词。为了让模型能够捕捉到各种潜在的动词化谓词, 我们使用 KELM (Agarwal 等人, 2021 年) 自动生成

的知识文本对齐数据来扩展任务。我们不包括其他远距离监督数据集（如 T-Rex ([El-sahar et al.](#))

对于 KELM 数据，由于它基于完整的 Wikidata 模式（其中包含太多关系，无法一一列举），我们为**关系提取**和**知识槽填充**任务创建了两个 KELM 专用提示：

(关系提取, 提示 1, 仅限 KELM)

```
{# kelm #}
你能从这句话中找出与整个维基数据属性有关的所有知识三元组吗? 请以"( X ; Y ; Z )"的形式列出它们:

{{text}} => ||| {{format_triple(relations, "") | join(" ")}}
```

(填写知识槽, 提示 1, 仅限 KELM)

```
{# kelm #}
给定实体"{{entities[entity_idx]}}"在上下文中标有"["和"]":

{{文本}}

请列出与之相关的所有三元组 (如果没有答案, 则不生成): ||| {% for r in relations
%}{% if r['head'][0] == entities[ entity_idx] %}{{format_triple([r], "")
| join(" ")}}{% endif %}{% endfor
%}
```

C.2.4 命名实体识别

命名实体识别是一项从原始文本语料中识别命名实体并为其分配适当实体类型的任务。例如, 在 "1916 年, 通用汽车公司在底特律被重新命名为通用汽车公司" 这句话中, 通用汽车公司可能属于组织实体类型。我们根据命名实体识别数据集 CoNLL03 ([Sang 和 Meulder, 2003 年](#))、OntoNotes 5.0 ([Pradhan 等人, 2013 年](#)) 和 GENIA ([Ohta 等人, 2002 年](#)) 设计了两种不同类型的任务。我们还包括来自联合实体和关系数据集的命名实体识别子任务。

- **命名实体识别:** 给定一定的可能实体类型列表 (如地点、个人、组织), 从提供的文本内容中提取所有相关实体。
- **实体类型:** 实体类型是命名实体识别的重要衍生任务之一。其目的是对实体提及 (无实体类型) 的正确类型进行分类, 通常作为后处理附加到实体提及提取中。

(命名实体识别, 提示 0)

给定以下实体类型列表:

```
Z = {{shuffle(allowed_types) | join(", ")}}
```

请从左至右提取句子中提到的所有实体, 形式为 "(X ; 实例 ; Z)"。

```
{{text}} => ||| {% for entity, type in zip(entities, entity_types) %}
{{实体}}; {{类型}}的实例{% endfor %}
```

(实体打字, 提示 0)

提取句子中提到的所有实体, 实体类型为 "{{ allowed_types[type_idx] }}" , 形式为 "(X ; instance of ; {{ allowed_types[type_idx] }})"。

```
{{text}} => ||| {% for entity, type in zip(entities, entity_types) %}{%
if type == allowed_types[type_idx] %}({{entity}} ; instance of ; {{type
}}){% endif %}{% endfor %}
```

作为会议论文发表于 2023 年国际比较文学
和历史研究国际会议 (ICLR 2023) 。

(实体打字, 提示 1)

列出以下段落中出现的所有"`{{allowed_types[type_idx]}}`"实体, 用"`|`"连接:

```
{{text}} => ||| {{filter_type(zip(entities, entity_types), allowed_types
[type_idx]) | join(" | ")}}
```

(实体打字, 提示 2)

```
{% if entity_types.len () > 0 %}
```

基于潜在实体类型列表, 忽略其顺序:

```
- {{shuffle(allowed_types) | join("\n- ")}}
```

的实体"`{{entities[entity_idx]}}`"标上"`"和"`":

```
{{文本}}
```

```
属于 ||| {{entity_types[entity_idx]}}
```

```
{% endif %}
```

C.2.5 关系分类

关系分类是信息提取中的一项基本任务, 它可以从候选列表中识别出两个给定实体之间的关系。这个问题由来已久, 因为它受制于高昂的数据标注成本, 因为对知识密集型任务进行人工标注需要受过教育的标注者, 而这些标注者的收费很高。事实上, 关系外的数据创建方法依赖于远距离监督, 即自动将知识库中现有的知识三元组与文本内容对齐, 并假定这种对齐在特定条件下是正确的。在这里, 我们只包括 TacRED (Zhang 等人, 2017 年) 数据集, 并在此基础上创建了几个不同的任务。

- **关系分类:** 最传统的任务表述方式。从文本中给出两个实体, 并从候选列表中对它们的关系进行分类。形式可以是直接回答关系, 也可以是三重形式 (类似于关系提取)。
- **知识槽填充:** 将任务改为给定头部实体和关系, 以识别输入文本中是否存在尾部实体。如果不存在, 则不生成任何内容。
- **是或否问题:** 将问题转化为类似于自然语言推理的任务。例如, 给定句子 "该系列主要讲述海滩男孩乐队创始人布莱恩-威尔逊的女儿卡尼-威尔逊的生平", 要求模型通过回答 "是" 或 "否" 来判断卡尼-威尔逊、父亲、布莱恩-威尔逊这三者的正确性。

(关系分类, 提示 0)

```
{% if entity_types.  
鉴于以下关系类别
```

```
- {{shuffle(allowed_relations.values()) | join("/n- ")}}
```

预测下面句子中"{{relations[0]['head']}}"和"{{relations [0]['tail']}}"之间的
关系:

{{文本}}

关系应为 : || {{allowed_relations[relations[0]['relation ']]}}
{% endif %}

(关系分类, 提示 1)

1. (关系提取) 以 "(X ; Y ; Z)" 的形式回答实体之间的关系:

{{文本}}

{{relations[0]['head']}} 和 {{relations[0]['tail']}} 之间的关系是: ||| ({{relations[0]['head']}} ; {{allowed_relations[关系[0]['关系']}} ; {{relations[0]['tail']}})

(填充知识槽, 提示 0)

根据下面提供的句子, 推断问题中缺少的论据:

{{文本}}

问题{{relations[0]['head']}} "是什么/谁/在哪里? ({{allowed_relations[relations[0]['relation']]}})?

请回答: || {{relations[0]['tail']}}

C.2.6 语义角色标签

语义角色标注是一项历史悠久的信息任务, 旨在识别句子中与给定谓词相关的语义论据。

例如, 在句子 "格兰特在 IBM 工作了 21 年, 担任过多个行政职务" 和谓词 "受雇" 中, 语义角色标注可以识别出格兰特是主语, IBM 是次宾语。

我们根据语义角色标签数据集 CoNLL05 (Carreras 和 Màrquez, 2005 年)、CoNLL12 (Pradhan 等人, 2013 年) 和 PropBank (Kingsbury 和 Palmer) 创建了两个不同的任务。

- **语义角色标签:** 传统的任务形式, 即在文本中标注一个动词 (即谓词), 并要求模型生成相关的语义角色。
- **语义角色填充:** 给定一个动词和一个潜在的语义角色, 要求模型判断句子中是否存在该角色并生成它。
- **谓语识别:** 给定一个句段及其相应的语义角色, 识别它与哪个动词相关。

(语义角色标签, 提示 0)

根据下面句子中标有 "[和]" 的目标动词 "{{动词}}", 找出它的

"{{allowed_types[type_idx]}}":

(语义角色填充, 提示 0)

给定以下参数类型列表:

Z = {{allowed_types | join(", ")}}

从左到右找出下面句子中提到的与动词 "{{动词}}" 有关的所有分论点, 形式为 "(X ; instance of ; Z)".

{{text}} => ||| {% for entity, type in zip(entities, entity_types) %} ({{实体}} ; 参数类型; {{类型}}) {% endfor %}

作为会议论文发表于 2023 年国际比较文学
和历史研究国际会议 (ICLR 2023) 。

(谓语句识别, 提示 0)

期末考试

1. 根据"{{entities[entity_idx]}}"是"{{entity_types[entity_idx]}}"的事实, 下面句子中的哪个动词应该是
它与什么有关?

{{文本}}

请回答: || {{verb}}

C.3 GPT-3、BLOOM-176B 和 OPT-175B 的成果来源

这里我们将介绍 GPT-3、BLOOM-176B 和 OPT-175B 的结果来源。我们可能比较的其他 LLM 大多是完全闭源的; 因此, 它们的结果都来自现有的预印本、出版物或存储在 BIG-bench 存储库¹⁰ 中的结果。

对于 GPT-3, 本文中的大部分结果如果没有具体说明, 则来自现有文献, 其余结果则通过我们自己申请的 OpenAI Danvici API 获取, 并明确提及。至于 BLOOM-176B 和 OPT-175B, 如果没有具体注释, 其结果如下:

- 摘自 OPT 论文 (Zhang 等人, 2022 年)。
- 摘自 EAI-Eval BigScience Arch&Scale - Google Sheet¹¹。
- 摘自 Huggingface Datasets¹² 中的 BigScience 评估结果库。

具体来说, 我们无法自行评估 OPT-175B 的情况, 因为我们仍未正式获得该检查点, 尽管我们在过去几个月中已递交了几份申请。

C.4 桩测试装置评估

Pile evaluation (Gao 等人, 2020 年) 是一个综合性语言建模基准, 最初包括 22 个不同领域的文本数据集。我们报告的是 18 个数据集的部分结果, 以及之前报告的基线结果 (Lieber 等人, 2021 年)。与传统的语言建模基准不同, Pile 评估报告的是 BPB (每字节比特数) perplexity, 以避免不同词汇量的模型之间的不匹配比较。因为一般来说, 如果不受限制, 词汇量较大的语言模型在复杂度比较中会更有优势。在评估过程中, 我们严格按照 (Gao 等, 2020) 中的设置, 利用 [gMASK] 和双向注意的上下文长度 1,024, 以及其余 1024 个标记,

表 13: GLM-130B 及其类似型号

以自回归的方式计算 BPB。加权平均 BPB 是根据 Pile 训练集中每个共享数据集的比例计算得出的 (Gao 等, 2020 年)。

	stackexchange	0.655	0.773	0.611
	nih_exporter	0.590	0.612	0.614
	pubmed_abstracts	0.587	0.625	0.610
	uspto_backgrounds	0.537	0.566	0.537
作为会议论文发表于 2023 年国际比较文学	pubmed_central	0.579	0.690	0.510
和历史研究国际会议 (ICLR 2023)。	freelaw	0.514	0.612	0.499
	github	0.358	0.645	0.329
	电子邮箱	0.621	0.958	0.604
	字幕	0.825	0.815	0.746
LLMs 在 Pile 测试集上的 BPB 结果。	加权平均值	0.650	0.742	0.634

Pile 测试集的详细指标见表 13。我们注意到，与 GPT-3 相比，GLM-130B 在 phil_papers 和 pile_cc 上的性能明显较弱，这可能是由于 GLM-130B 的双语天然性以及缺乏更多样化和高质量的私人收集语料所致。

¹⁰<https://github.com/google/BIG-bench>

¹¹<https://docs.google.com/spreadsheets/d/1CI8Q9RCblLRzUOPJ6ViqBmo284-8oj1uQ-CmaEuhuv0>

¹²https://huggingface.co/datasets/bigscience/evaluation-results/tree/main/bloom/bloomzeval/transformers/evaluation_val

C.5 BIG-BENCH-LITE 评估

最近的研究 (Wei 等人, 2022c; Wang 等人, 2022c) 再次证明, LLMs 有能力完成常规语言任务之外的推理。作为回应, BIG-bench (Srivastava et al. 出于经济方面的考虑, 我们在原 150 个任务的 BIG-bench 的官方子集, 即包含 24 个任务的 BIG-bench-lite 上对 GLM-130B 进行了评估。这些任务可分为两类: 一类是基于带答案选项的多选题回答, 另一类是不带选项的直接生成。对于第一类任务, 我们评估每个选项全部内容的概率, 并挑选最大的一个作为答案; 对于第二类任务, 我们使用贪婪去编码生成答案。在 BIG-bench 中进行的所有评估都基于 [MASK], 因为这里的答案通常都是短小的文本。在 24 个 BIG-bench-lite 数据集 (Srivastava et al.

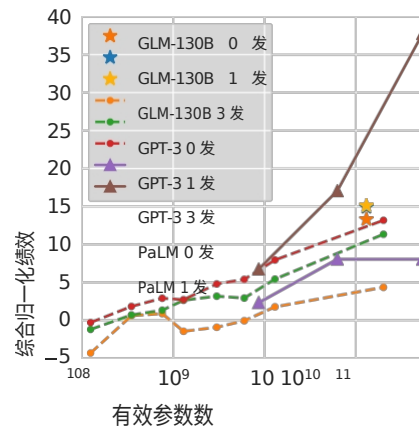


图 16: BIG-bench-lite (24 项任务) 的全部评估范围。

图 16. 我们只是采用了 BIG-bench 的原始提示, 并使用官方实现来生成引物示例, 以进行少量评估并计算最终分数。

C.6 毫米路评估

57 个 MMLU (Hendrycks 等人, 2021 年) 数据集 GLM-130B 和 BLOOM 176B 的所有结果见表 15。在第 5.2 节中, 我们报告了 GLM-130B、GPT-3 175B 和 BLOOM 176B 的加权平均准确率 (即每个样本的平均准确率, 而不是按学科)。

下面是一个单次引物的提示示例。我们预测下一个标记 ['A'、'B'、'C'、'D'] 的概率, 并选择概率最大的一个作为答案。

(MLU 1shot 示例)

以下是有关哲学的多项选择题。

根据霍尔巴赫的观点, 人们总是根据 _____。

(A) 自由选择 (B) 灵魂的支配 (C) 必要的自然规律 (D) 不确定的意志

答案: (C) 必要的自然规律

伊壁鸠鲁认为, 哲学就是

(A) 不适合年轻人。 (B) 不适合老年人。 (C) 重要, 但不愉快。 (D) 以上都不是。

答案: (

C.7 中文理解能力评估

在此, 我们将详细阐述我们在 CLUE (Xu 等人, 2020 年) 和 FewCLUE (Xu 等人, 2021

年) 评估中使用的提示语。在中文数据集上, 提示遇到了一些挑战, 因为中文文本是由单字而不是单词组织的, 这在很多情况下会导致动词长度不等。尽管针对特定数据集的校准 (Wang 等人, 2021 年; Wu 等人, 2021 年) 有助于缓解这一问题, 但过于具体的技术在实施时可能会很复杂。本文的评估利用 GLM-130B 的独特功能, 采用了一种更易于解决的方法。由于 GLM-130B 是一个具有英语 MIP 的双语 LLM, 我们采用了 (Bach 等人, 2022 年) 类似任务中的英语提示和口头化器来进行中文数据集评估, 并发现这些策略相当有效。在评价指标方面, 除了 DRCD 和 CMRC2018 两个问题解答数据集报告了 EM 外, 其他数据集都报告了准确率。

C.8 自然语言生成

这里的自然语言生成或条件自然语言生成是指需要根据给定信息生成文本的任务，如表格和文档。我们对 GLM-130B 的数据到文本和摘要任务进行了评估。数据集包括 WebNLG 2020 (Cas-tro Ferreira 等人, 2020 年)、Clean E2E NLG (Dušek 等人, 2019 年) 和来自 GEM 生成基准 (Gehrmann 等人, 2021 年) 的 WikiLingua (Scialom 等人, 2020 年)。我们在测试集中选择了完整的 WebNLG 2020 和 Clean E2E NLG，并按照 (Chowdhery et al.) 按照 PaLM 的设置，总结任务使用的提示语是 "总结以下文章："，数据到文本任务使用的提示语是 "口头表达："。E2E 是一个例外，我们使用 GLM-130B 和 GPT-3 175B 的 promptsource (Davinci) 中提供的提示 "generate-gramatically-correct-text from "来处理数据。所有评估都是一次性的，演示样本是从训练集中随机抽取的。我们报告了 ROUGE-2、ROUGE-L (Lin, 2004 年) 和 BLEURT-20 (Pu 等, 2021 年) 的 F-measure。我们将我们的模型与 LaMDA、GPT-3 175B (Davinci) 和 PaLM 进行了比较，其中 LaMDA 和 PaLM 的结果由 (Chowdhery 等人, 2022 年) 报告，我们通过 OpenAI API 对 GPT-3 175B (Davinci) 进行了评估。¹³

结果见表 16。结果表明，在所有任务中，GLM-130B 的性能都优于 LaMDA 和 GPT-3 (Davinci)。在数据到文本任务中，GLM-130B 的表现略逊于 PaLM-540B，而在摘要任务中，GLM-130B 的 ROUGE 结果甚至更高。我们还将 GLM-130B 消减为单向，以证明双向注意力的优势。在所有三个数据集中，单向 GLM-130B 的表现都不如 GPT-3 175B，但当它转向双向注意时，表现立即得到提升，在少数情况下，GLM-130B 甚至可以与 PaLM- 540B 相媲美。这表明，对提供的上下文（即前缀）进行双向关注也有利于文本生成任务。

表 16：1 次 GEM 英语自然语言生成任务（WebNLG、E2E 和 WikiLingua）。我们比较了 GLM-130B 的两个版本（uni：单向注意；bi：双向注意），结果表明双向注意也能提高条件生成的性能。

任务	数据集	度量	LaMDA 137B	GPT-3 175B (还又四)	GLM-130B		PaLM-540B
					单轴	偏颇	
数据 至 文 本	网络版	胭脂-2	30.5	29.9	25.3	<u>38.5</u>	44.4
		红酒-L	-	41.2	36.7	<u>49.3</u>	53.8
		BLEURT-20	-	59.0	53.2	<u>67.7</u>	73.9
	E2E	胭脂-2	29.2	30.3	30.9	<u>33.9</u>	35.2
		ROUGE-L	-	39.2	40.0	<u>42.6</u>	43.9
		BLEURT-20	-	64.5	65.0	<u>68.1</u>	69.7
摘要	维基语言	胭脂-2	5.4	7.2	5.8	10.4	<u>9.9</u>
		ROUGE-L	-	18.9	16.4	23.4	<u>20.6</u>
		BLEURT-20	-	41.2	39.4	<u>45.0</u>	47.7

(E2E 示例, 无示范样本)

Aleksandr_Prudnikov , height , 185.0
(centimetres).FC_Spartak_Moscow , 场地 , Otkrytiye_Arena.
Aleksandr_Prudnikov , club , FC_Spartak_Moscow.口述:

Groundtruth: 身高 185 厘米的 Aleksandr Prudnikov 曾效力于莫斯科 Otkrytiye 竞技场的斯巴达克俱乐部。

GPT-3 175B (达文西) : Aleksandr Prudnikov 是俄罗斯莫斯科斯巴达克足球俱乐部的一名中场球员。
。

GLM-130B: Aleksandr Prudnikov 身高 185.0 厘米, 效力于莫斯科斯巴达克俱乐部。

¹³我们使用 <https://github.com/google-research/google-research/tree/master/rouge> 上的 ROUGE 实现和 <https://github.com/google-research/google-research/tree/master/rouge> 上的 BLEURT-20 实现, 其检查点见 <https://storage.googleapis.com/bleurt-oss-21/BLEURT-20.zip>

(E2E 示例, 无示范样本)

将以下所有数据合并成简明扼要、语法正确的文本:

name : Blue Spice
eatType : 咖啡店 area :
riverside

地面实况: 在河边, 有一家名为 "蓝色香料" 的咖啡店。

GPT-3 175B (达文西): Blue Spice 是一家河畔咖啡店, 位于河街 (River Street) 和河岸街 (Riverbank Street) 的拐角处。

GLM-130B: 河畔地区有一家供应咖啡的咖啡店, 名叫 "蓝色香料"。

(WikiLingua 示例, 无示范样本)

您的大多数客户都会在网上搜索您, 因此拥有一个用户友好型网站至关重要。您的网站至少应包括有关您的业务和您在搬家行业的历史、报价流程详情、联系信息以及您所提供服务的说明。如果可能, 允许客户在线安排报价、查看可用性或阅读其他客户的评价。最简单的方法之一是

您的业务就是帮助您已经认识的人搬家。您可以留意朋友在社交媒体上发布的任何有关搬家的公告。一旦您为朋友提供了良好的服务, 他们很可能会向其他人推荐您。为了宣传您的业务, 可以印制一些专业的宣传材料, 并在社区内分发。

您可以在公共活动中分发名片, 将其夹在当地的公告栏中, 甚至将其印制在通讯录、年鉴以及其他当地印刷媒体。可以邮寄传单、在公共场所张贴传单, 或将传单分发给有可能将客户介绍到以下地点的企业

如家具店。确保你有一个专业的、可识别的徽标, 并在所有营销材料中保持一致。另一种让企业名称广为人知的方法是让自己引人注目。无论是在当地活动中与合作伙伴合作、

无论是志愿服务, 还是利用车辆开展广告活动, 知名度都是推动业务发展的关键。在以下领域与有影响力的人士建立关系

您所在的社区。房地产经纪人是介绍搬家公司的重要来源, 当地家具店老板或大型公寓大楼的办公室人员也是如此。您可以使用 `Craigslist` 等目录网站免费向当地社区的人宣传您的服务。社交媒体也是传播企业信息的好方法。

根据您的预算和目标市场的不同, 有很多广告选择。可以考虑 `PPC` 广告、电视和广播广告、报纸广告、直邮传单或会员制等方式。

转介服务。要想发展业务, 最好的办法就是为客户提供优质服务。请务必做到准时、友好、尊重客户的物品, 以及

提供准确的报价。一定要让您的满意客户在网上发表对您的评价。新客户会看到这些评论, 如果他们看到别人对您的评论, 就会更加相信您是一家信誉良好的公司。

有良好的体验。在某些时候, 你难免会遇到不满意的顾客, 但你要尽最大努力解决问题, 让顾客满意。你最不希望看到的就是网上流传着对你企业的负面评价! 总结

这篇文章:

真相: 创建一个网站。从朋友开始。分发名片和传单。建立宣传网络。寻找免费广告机会。考虑付费广告。赢得良好声誉。

GPT-3 175B (达文西): 在网上推销自己。寻找机会在线下推销自己的服务。

作为会议论文发表于 2023 年国际比较文学

和历史研究国际会议 (ICLR 2023) 。

GLM-130B: 拥有用户友好型网站。通过社交媒体推广业务。与社区中有影响力的人建立关系。使用 Craigslist 等目录网站宣传您的服务。

表 17: Winograd 风格任务评估 (Winogender 和 Winograd273)。所有分数均为准确度。K 指射击次数。*PaLM 540B 没有报告精确的 0 次 Winogender 结果, 因此我们必须从其绘制的图表中估算出一个值。

	K	GPT-3 (达文西)	OPT 175B	生机勃勃 176B	PaLM 540B	栗鼠	地鼠 280B	GLM-130B
Winogender	0	64.2	54.8	49.1	75.0*	78.3	71.4	79.7
	1	62.6	-	53.1	79.4	-	-	80.7
Winograd273	0	88.3	52.9	49.1	90.1	-	-	84.3

表 18: 闭卷答题 (自然问题、StrategyQA)。

	GPT-3 (达文西)	BLOOM 176B	PaLM 540B	栗鼠	地鼠 280B	GLM-130B
自然问题 (EM)	14.6	13.1	21.2	16.6	10.1	11.7
StrategyQA (Acc)	52.3	49.8	64.0	-	-	60.6

表 19: 常识推理 (常识 QA, MC-TACO)。K 指拍摄次数。

	K	GPT-3 (达文西)	OPT 175B	BLOOM 176B	GLM-130B
常识性 QA (Acc)	0	57.2	-	42.8	61.6
	1	61.2	-	-	62.2
MC-TACO (EM)	0	-	12.4	13.1	13.6

C.9 维诺格拉德式任务

我们包括对 Winograd 风格任务的评估, 该任务源自经典的 Winograd 模式挑战赛 (Levesque 等人, 2012 年), 旨在测试机器在模棱两可的语境中的核心参照解析能力。由于在 MIP 中, 我们已经包含了 Winogrande (Sakaguchi 等人, 2021 年) 和 SuperGLUE WSC (Wang 等人, 2019 年), 因此我们在此测试 Winogender (Rudinger 等人, 2018 年) 和 Winograd273 (Levesque 等人, 2012 年)。对于 Winogender, GPT-3 的结果来自 OpenAI API, 而 BLOOM 的单发结果则由我们自己评估。对于 Winograd273, 由于现有研究 (Brown 等人, 2020; Chowdhery 等人, 2022) 表明单次学习几乎没有带来任何改进, 因此我们只测试了零次学习的结果。另一个值得注意的问题是, 尽管 GPT 风格的模型 (如 GPT-3、PaLM) 采用了 (Radford et al.

结果见表 17。在 Winogender 上, GLM-130B 在所有评估的 LLM 中表现最佳, 而在 Winograd273 上则略逊于 GPT-3 和 PaLM。

C.10 闭卷答题

闭卷答题 (CBQA) (Roberts 等人, 2020 年) 是一项广泛采用的任务, 用于评估语言模型对事实知识的记忆, 与传统的 "开卷" 评估相反。由于我们在 MIP 训练中加入了 TriviaQA (Joshi 等人, 2017 年) 和 WebQuestions (Berant 等人, 2013 年), 因此我们在此选择 Natural Questions (Kwiatkowski 等人, 2019 年) 和 StrategyQA (Geva 等人, 2021 年) 作为 CBQA 的评估数据集。

结果见表 18。GLM-130B 在自然问题上的表现相对较差, 而在 StrategyQA 上表现良好。我们推测, GLM-130B 在自然问题上的表现不佳, 可能是由于对英语语料的拟合不足, 因为它大致只查看了

200B 英语词块，因此不能很好地记忆详细知识。由于 CBQA 似乎是一项特别强调记忆的任务，正如 Chinchilla (Hoffmann 等人, 2022 年) 的出色表现所表明的那样，我们认为在以后的充分训练中，GLM-130B 可以表现得更好。

C.11 常识推理

在此，我们对 GLM-130B 和其他一些 LLM 的常识推理能力进行了评估。由于我们已经将 PIQA (Bisk 等人, 2020 年)、ARC (Clark 等人, 2018 年) 和 OpenbookQA (Mihaylov 等人, 2018 年) 纳入了 MIP 训练，因此我们选择了另外两个广泛采用的常识推理数据集进行评估：常识 QA (Talmor 等人, 2019 年) 和 多选时态常识 (MC-TACO, Zhou 等人, 2019 年)。对于常识问答，我们通过 OpenAI Davinci API 测试了 GPT-3，通过 Huggingface 实现测试了 BLOOM-176B，并使用 promptsource (Bach 等人, 2022 年) 中的 "answer_given_question_without_options" 提示测试了 GLM-130B。对于 StrategyQA，我们采用了 (Zhou 等人, 2019 年) 中提供的 EM 计算方法。

结果如表 19 所示。我们可以看到，在所有已评估的 LLM 中，GLM-130B 在常识性 QA 和 MC-TACO 方面的表现都是最好的，这表明 GLM-130B 对常识性知识掌握得很好。由于附录 C.3 中描述的原因，OPT 的结果未包括在内。

C.12 固定标签数据集：自然语言推理案例研究

正如第 5 节所讨论的，由于使用了 MIP，我们在 GLM-130B 的评估中采用了相当严格的零/少镜头学习数据集选择标准。然而，这一标准大大减少了我们目前可以评估的数据集，特别是一些读者怀疑不在 MIP 查看过的固定标签数据集上进行评估的限制是否必要（例如自然语言推理 (NLI)），并建议我们可以在一个独立的章节中进行报告，以避免混淆。

坦率地说，在这种情况下，GLM-130B 的零/少次学习可能会很有优势。下面，我们以 NLI 为典型例子，展示 GLM-130B 在这些场景中的优势。我们将 6 个广泛使用的 NLI 数据集（GLM-130B 的 MIP 训练中没有包含这些数据集）作为基准。结果如表 20 所示，从表中可以看出，由于任务类型的不同，GLM-130B 的 "0-shot" 性能要好得多。

表 20：GLM-130B 在 6 个典型自然语言推理 (NLI) 数据集上的 "零次" 结果。

*免责声明：尽管我们从未见过这些数据集，但一些其他的 NLI 数据集已被输入。在 GLM-130B 的 MIP 中包含了这一功能，使其有别于现有的标准零镜头设置。

	BLOOM 176B	OPT 175B	GLM-130B*
qnli (有效, 5 个提示的中位数)	50.9	55.4	86.7
mnli (有效, 15 个提示的中位数)	35.5	36.0	85.7
mnli_mismatched (有效, 15 个提示的中位数)	35.5	36.0	84.6
wnli (有效, 5 个提示的中位数)	57.7	53.5	67.6
胶水/可乐 (有效, 5 个提示的中位数)	39.0	44.4	57.6

glue/mrpc (有效, 5 个提示的中位数)	31.6	44.6	87.3
---------------------------	------	------	------

C.13 超级月亮

我们还报告了在 SuperGLUE (Wang 等人, 2019 年) 基准测试中对 GLM-130B 的评估, 该基准测试包括 8 个不同的自然语言理解挑战。需要注意的是, 这些结果既不是零/少量结果, 也不是微调结果, 因为 8 个任务中有 7 个任务的训练集已与其他 67 个多任务数据集一起包含在 GLM-130B 的 MIP 训练中 (ReCoRD 除外); 但是, GLM-130B 也没有在其中任何一个任务上进行单独微调。因此, 这些结果并非用于与其他模型的相对比较, 而仅供读者参考 GLM-130B 的绝对能力。

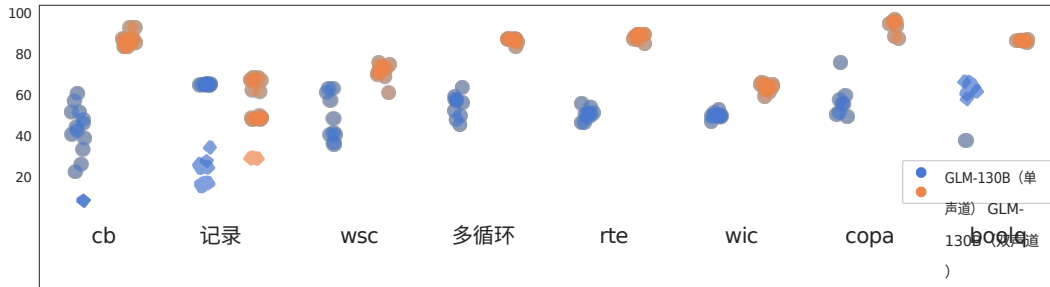


图 17: GLM-130B (uni 和 bi) 在 SuperGLUE 开发集上使用提示源 (Bach 等人, 2022 年) 提示和任务制定的未调整结果。免责声明: 注意到部分 SuperGLUE 训练集已被纳入 MIP 训练。我们在此报告结果, 仅供读者参考。

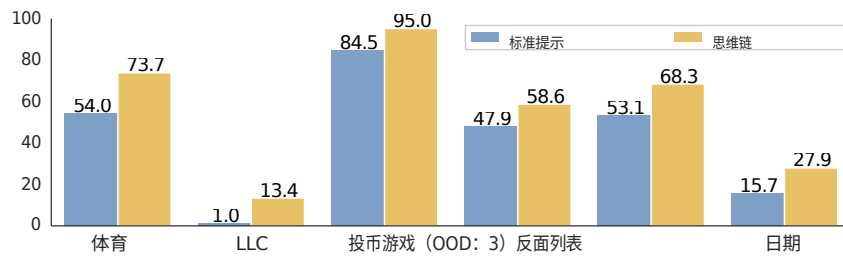


图 18: 与标准提示相比, 思维链提示也能提高 GLM-130B 在推理任务中的表现。

BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	GLM-130B	89.69
98.21	100	89.32	92.11	94.22	76.96	88.5			

表 21: 使用 P-tuning v2 (Liu 等人, 2022 年) 获得的 GLM-130B 在 SuperGLUE 数据集上的结果。除 MultiRC (F1a) 和 ReCoRD (F1) 外, 我们报告了所有数据集的精度指标。

结果如图 17 所示。我们对单向和双向 GLM-130B 进行了消减, 以证明 GLM 目标在提高 LLM 理解能力方面的作用。图中的每个点都是针对特定提示的结果, 其提示来自提示源 (Bach et al. 我们也采用了 promptsource 中的任务表述。我们可以看到, GLM (bi) 在所有任务中的方差都更小, 性能也更高。对于某些任务 (如 CB、MultiRC、RTE、COPA 和 BoolQ), GLM-130B 甚至可以达到 80% 以上的准确率。

我们还尝试在 SuperGLUE 数据集上对 GLM-130B 进行微调。然而, 当我们在下游任务中使用全参数微调时, 我们遇到了在单个历时内快速过拟合的问题。这导致在验证集上的性能不佳。为了解决这个问题, 我们探索了使用高效的参数微调方法, 这种方法只调整少量参数, 不易出现过拟合。在对几种方法进行实验后, 我们使用了 P-Tuning v2 (Liu 等人, 2022 年), 其结果与 GLM-130B 中的完全参数微调相当, 但调整的参数只占 0.1% 到 3%。表 21 列出了我们使用 P-Tuning v2 的实验结果。

C.14 思维链提示

我们按照 Wei 等人 (2022c) 中的设置，在**最后字母连接** (LLC)、**硬币翻转**、**反向列表**和来自 BIG-bench Srivastava 等人 (2022) 的两项任务中评估了思维链提示性能。结果如图 17 所示。我们发现思维链提示可以提高 GLM-130B 在符号推理和常识推理方面的表现。

日志缩放能力任务

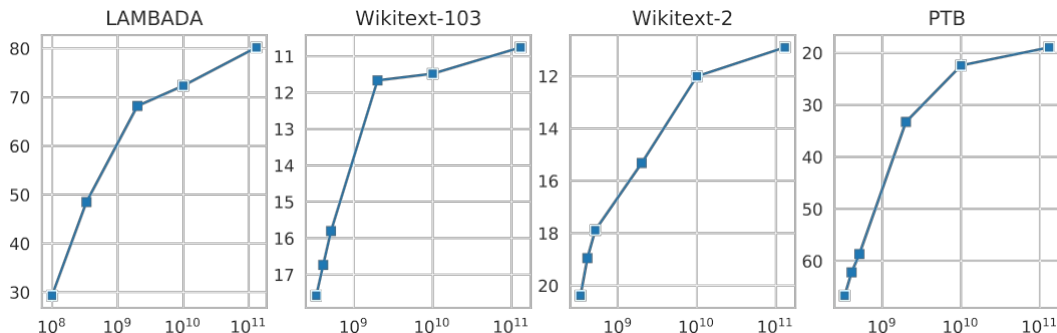


图 19: GLM-130B 的对数缩放能力任务。这些任务的性能随着 GLM 参数数量的增加而呈对数增长。大多数传统的 NLP 任务都属于相同的模式。

最后字母连接 (LLC)。该任务要求模型将姓名中单词的最后一个字母连接起来 (例如, "Elon Musk" -> "nk")。我们通过随机连接姓名普查数据¹⁴中排名前 1000 位的姓和名来生成全名。

抛硬币。这项任务要求模型回答一枚硬币从正面开始, 在人们掷硬币或不掷硬币后, 是否仍然是正面。(例如: "一枚硬币正面朝上。菲比抛硬币。奥斯瓦尔多没有抛硬币。硬币还是正面朝上吗?" -> "不是")。->"否")。此外, 我们还对查询示例中的人数多于上下文示例中的人数的情况进行了评估, 即分布外 (OOD) 设置。

颠倒列表。这项任务要求模型颠倒日常物品列表的顺序 (例如, "雪茄、雨伞、钥匙、口香糖、警报器"->"警报器、口香糖、钥匙、雨伞、雪茄")。我们从日常物品词汇表¹⁵中随机抽样生成列表。

体育。这项任务要求模型判断有关体育运动员的陈述的真实性 (例如, "若昂-穆蒂尼奥在全美橄榄球锦标赛中接住了挡拆传球"->"假")。

日期。这项任务要求模型根据给定的上下文推断数据 (例如, "2015 年还有 36 小时就要到来。从今天起一周后的日期是什么, 用 MM/DD/YYYY 表示?" -> "01/05/2015")。-> "01/05/2015")。

我们使用了与 Wei 等人 (2022c) 相同的示例和链。对于每项任务, 我们都尝试了两种不同的提示格式以及单向和双向注意机制, 并报告了最佳表现。第一种格式是 "问题: 语境回答: {目标}"。第二种是在第一种提示格式的示例前添加序号。结果如图 18 所示。

D GLM-130B 的规模和新兴能力

事实证明, 扩大预训练语言模型的规模可以不断提高下游任务的性能。他的新兴能力是较小规模无法预测的。为了说明这一点, 我们进行了大量实验来探索缩放特性和新兴能力。

按照先前的文献 (Wei 等人, 2022b)，我们根据观察结果将 NLP 任务分为两类。

- **对数缩放能力任务 (参见图 19)**：任务性能随模型参数的数量呈对数增长。典型的任务和数据集包括 LAMBADA、Wikitext- 103、Wikitext-2 和 Penn Tree Bank。
- **新兴能力任务 (参见图 20)**：只有当模型参数量达到一定临界值时，任务性能才会飙升。典型的任务和数据集包括：

¹⁴<https://namecensus.com>

¹⁵<https://www.vocabulary.com/lists/189583>