大语言模型实体标注规范

本规范针对大语言模型数据集构建过程中的实体标注流程制定，数据集来源为大语言模型模型介绍页（README.md）以及大语言模型相关论文，所包含的实体类别有模型名称、模型系列、开源情况、模型类型、参数量、支持语言、最大序列长度、开发机构八类。

1 模型名称 </name>

模型名称是大语言模型的基础内容，是大语言模型相关数据内容的最重要的、首先的关注点，主要是模型系列下的各个参数量版本模型，与模型系列的鲜明区别在于，要求有明确的模型参数量标识，后缀一定有参数量 b/B。

1.1 简单的模型名称

简单的模型名称即文本中明确给出“模型系列-参数量”的形式，包括但不限于微调后的模型系列，且无其他后缀。

例如：

<name>Baichuan-13b</name> 使用了 ALiBi 线性偏置技术，相对于 Rotary Embedding 计算量更小，对推理性能有显著提升。

The code for [recovering **<name>Chinese-Alpaca-2-13B</name>** weights from our released weight diff].

1.2 复杂的模型名称

复杂的模型名称即文本中未明确给出“模型系列-参数量”的形式，包括但不限于微调后的模型系列，有其他后缀，此时将后缀省去即可，且从开头保留模型系列内容，链接中的内容也要标注。

例如：

[2024/05/30] 发布**<name>Llama-3-Chinese-8B</name>**-Instruct-v3 版指令模型，相比 v1/v2 在下游任务上获得显著提升。

(https://byfintech.medium.com/beginners-guide-to-fingpt-training-with-lora-**<name>chatglm2-6b</name>**-9eb5ace7fe99).

https://huggingface.co/chitanda/**<name>llama-Panda-13B</name>**-zh-wudao-chat-delta

1.3 特别要注意的模型名称

假如说有一个模型名称为“机构名/领域-模型系列-参数量”，将机构名摘出来单独标注，领域暂不做处理，这种情况出现的较少，但需要特别注意。

例如：

发布**<name>Llama-3-Chinese-8B</name>**-Instruct-v2 版指令模型，直接采用 500 万条指令数据在 [**<in>Meta</in>-<name>Llama-3-8B</name>**-Instruct](https://huggingface.co/meta-llama/**<in>Meta</in>-<name>Llama-3-8B</name>**-Instruct) 上进行精调。

最新 Code-**<name>Panda-13B</name>**-Python、Legal-**<name>Panda-13B</name>**-Chat

2 模型系列 </series>

模型系列是大语言模型的系列名称，一般是大语言模型相关数据内容介绍最多的部分，主要是模型系列发布的名称或者是论文中提到的模型系列，与模型名称的鲜明区别在于，要求不带有明确的模型参数量标识，尽量精简，带参数量标识的一律归为模型名称。

2.1 简单的模型系列

简单的模型系列即文本中明确直接给出“模型系列”的形式，包括但不限于后缀的版本信息，需要将其一同标注在内。

例如：

针对**<series>OpenLLaMA</series>**，在训练过程中我们发现其需要接近两倍的训练时长，且最后确认不是计算节点通信的问题，我们怀疑是**<series>OpenLLaMA</series>**的模型精度问题导致了训练降速。

[**<series>BianQue2</series>**](https://huggingface.co/scutcyr/**<series>BianQue-2</series>**)(**<series>BianQue-2.0</series>**)

<series>巧板</series>大模型是一个**<params>7B</params>**规模的大语言模型。

2.2 复杂的模型系列

复杂的模型系列即文本中明确直接给出“模型系列-模型系列”的形式，包括但不限于后缀的训练基座信息，需要分开标注，版本信息仍保留在内。

例如：

达摩院的同学发布[Video-**<series>llama</series>**-**<series>BiLLa</series>**]

<series>星辰语义大模型</series>-**<series>TeleChat</series>**

我们使用[**<series>Alpaca</series>**-**<series>GPT4</series>**]进行后续实验

2.3 特别要注意的模型系列

假如说有一个模型系列为“模型系列-模型系列-参数量”，则将其整体统一标注为模型名称，不再单独得将其标注为模型系列，需要特别注意。

例如：

[Huggingface Ziya-visual Space](https://huggingface.co/spaces/**<in>IDEA-CCNL</in>**/**<name>Ziya-BLIP2-14B</name>**-Visual-v1-Demo)

We release **<name>YuLan-LLaMA-2-13B</name>** and **<name>YuLan-Chat-2-13B</name>**.

3 开源情况 </status>

开源情况是指在文本描述中大语言模型的参数信息是否开源，或者是否公开发表出模型的权重文件以及模型配置信息，这里统一按照开源“open-source”，闭源“closed-source”来进行开源情况实体标注。

例如：

Our **<name>WizardMath-70B</name>**-V1.0 model slightly outperforms some **<status>closed-source</status>** LLMs on the GSM8K.

<series>Baize</series> is an **<status>Open-Source</status>** chat model trained with [LoRA].

4 模型类型 </type>

模型类型是指文本和多模态两类大语言模型，其中多模态大语言模型又分为语言大语言模型、图像大语言模型、视频大语言模型等等，这里仅标注上述四个类型，以及四个类型下的交叉类型“开源情况-开源情况”。

例如：

<status>开源</status>[**<series>BELLE-VL</series>**](https://huggingface.co/**<series>BELLE-2</series>**/**<series>BELLE-VL</series>**)**<type>多模态</type>**大语言模型

<type>语音-文本</type> **<type>视觉-文本</type>** **<type>视觉-语言</type>**

Introducing **<series>DeepSeek-VL</series>**, an **<status>open-source</status>** **<type>Vision-Language</type>** (VL) Model designed for real-world vision and language understanding applications. **<series>DeepSeek-VL</series>** possesses **<type>multimodal</type>** understanding capabilities

新增**<type>语音-语言</type>****<type>跨模态</type>**模型**<series>ERNIE</series>**-SAT[正式**<status>开源</status>**]

5 参数量 </params>

参数量是指大语言模型的参数数量，以“b/B”为结尾，值得注意的是，没有模型系列的前缀，即不包含在模型名称标签内的参数量信息。

例如：

<series>LLaVA-NeXT</series> (Stronger) models are released, stronger LMM with support of **<series>LLama-3</series>** (**<params>8B</params>**) and **<series>Qwen-1.5</series>** (**<params>72B</params>**/**<params>110B</params>**)

<params>7B</params>("Model Worker: **<name>llava-v1.5-7B</name>** Port:40000")

| **<params>7B</params>** | [**<name>MAmmoTH-7B</name>**](https://huggingface.co/**<in>TIGER Lab</in>**/**<name>MAmmoTH-7B</name>**) |

6 支持语言 </lang>

支持语言是指大语言模型所使用的训练文本语言，主要是中文、英文和双语，以及少数的小语种语言标注信息。

例如：

``` **<series>明医 (MING)</series>**：**<lang>中文</lang>**医疗问诊大模型

MOSS 是一个支持**<lang>中英双语</lang>**的**<status>开源</status>**对话语言模型

**<series>PolyLM</series>** is proficient in the major non-English languages spoken worldwide, such as **<lang>Spanish</lang>**, **<lang>Russian</lang>**, **<lang>Arabic</lang>**, **<lang>Japanese</lang>**, **<lang>Korean</lang>**, **<lang>Thai</lang>**, **<lang>Indonesian</lang>**, and **<lang>Chinese</lang>** etc.

## 7 最大序列长度 </contextl>

最大序列长度是指大语言模型允许输入的最大序列长度或者是多模态大语言模型允许输入的最大图像分辨率，数目主要是 2 的次方数，像 256/512/1024/2048/2k/4k...，值得注意的是，像部分模型名称后会接一个最大序列长度，这时需要将这里的模型名称和最大序列长度分别进行标注。
```

例如：

标准版模型支持**<contextl>4K</contextl>**上下文长度，长上下文版模型支持 **<contextl>16K</contextl>**、**<contextl>64K</contextl>**上下文长度。

分辨率对上述某几个评测非常重要，大部分**<contextl>224 分辨率</contextl>**的**<status>开源</status>**LVLM 模型无法完成以上评测。

<name>AquilaChat2-34B</name> 和 **<name>AquilaChat2-70B</name>**-Expr，长文本对话模型**<name>AquilaChat2-7B</name>-<contextl>16k</contextl>** 和 **<name>AquilaChat2-34B</name>-<contextl>16k</contextl>**，您可以通过点击下方图标进入下载界面：

8 开发机构 </in>

开发机构是指大语言模型的发布大学或者开发机构，主要标注常见的一些机构，一般在模型描述页的开头和最后，以及论文的介绍部分会直接给出，值得注意的是，“机构名-模型系列-参数量”，将机构名摘出来单独标注。

例如：

You can also access the ollama service via its **<in>OpenAI</in>**-compatible API.

最简单的使用**<series>Qwen</series>**模型 API 服务的方法就是通过 DashScope（**<in>阿里云</in>**灵积 API 模型服务）

基于主动健康的主动性、预防性、精确性、个性化、共建共享、自律性六大特征，**<in>华南理工大学未来技术学院</in>-<in>广东省数字孪生人重点实验室</in><status>开源</status>**了中文领域生活空间主动健康大模型基座。

发布**<name>Llama-3-Chinese-8B</name>**-Instruct-v2 版指令模型，直接采用 500 万条指令数据在 [**<in>Meta</in>-<name>Llama-3-8B</name>**-Instruct]上进行精调。