CCL24-Eval任务2系统报告：基于样本设计工程和大模型微调的中文意合图语义解析*

司函[1]，罗智勇[1†]
[1]北京语言大学 信息科学学院
202221198690@stu.blcu.edu.cn，luo_zy@blcu.edu.cn

摘要

本文介绍了我们在第二十三届中国计算语言学大会中文意合图语义解析评测中提交的参赛系统。中文意合图（Chinese Parataxis Graph，CPG）是以事件为中心的语义表征图，可以对不同层级的语言单元作一贯式表示，是一种通用性与扩展性兼具的语义表征方法。鉴于大语言模型在语义解析任务中的优越性能，我们对Llama3-Chinese-8B-Instruct模型进行了LoRA微调，使其能够生成结构化的意合图表征三元组，并采用了样本设计工程（Sample Design Engineering，SDE）技巧进行微调样本的设计。此外，我们还对不同标签进行了分类微调，探究大模型在不同语义标签预测能力上的差异。最终，我们的参赛系统在任务发布的评测集上F1值达到0.6461，在本次评测任务中获得了第三名的成绩。

关键词： 意合图；语义解析；大语言模型；样本设计工程

System Report for CCL24-Eval Task 2: The Chinese Parataxis Graph Parsing Based on Sample Design Engineering and Fine-Tuning Large Language Model

Han Si[1]，Zhiyong Luo[1*]
[1]School of Information Science, Beijing Language and Culture University
202221198690@stu.blcu.edu.cn，luo_zy@blcu.edu.cn

Abstract

This paper introduces the system we submitted in the shared task of Chinese Parataxis Graph (CPG) Parsing at the Twenty-three Chinese National Conference on Computational Linguistics. The Chinese Parataxis Graph is an event-centered semantic representation graph that provides a consistent representation of language units at different levels, offering a versatile and extensible method of semantic representation. Considering the superior performance of large language models in semantic parsing tasks, we fine-tuned the Llama3-Chinese-8B-Instruct model using LoRA to enable it to generate structured parataxis graph representation triples. We also employed the Sample Design Engineering (SDE) technique for the design of fine-tuning samples. Furthermore, we conducted classification fine-tuning for different labels to explore the model's performance in predicting various semantic labels. Ultimately, our system achieved a F1 score of 0.6461 on the evaluation set provided by the task organizers, securing the third place in this evaluation task.

* 基金项目：国家自然科学基金（62076037）
† 通讯作者

第二十三届中国计算语言学大会论文集，第72页-第79页，太原，中国，2024年7月25日至28日。
卷3：评测报告

Keywords: Chinese Parataxis Graph , Semantic Parsing , Large Language Model , Sample Design Engineering

1 引言

语义解析是指将自然语言文本转化为结构化的语义表示，使得计算机能够理解和执行人类的指令。语义解析是自然语言处理领域亟待突破的瓶颈，精准把握自然语言语义需要准确且完备的语义表示方法。英文语义表示的研究发展较早，最具典型的就是抽象语义表示(Abstract Meaning Representation，AMR) (Banarescu, 2013)，它是句子级语义表示方法的一种，它将句子中的事件、状态、属性等内容抽象为语义概念，通过图的方式表示不同概念之间的语义关系，图中的边则表示不同概念节点之间存在的语义关系。随着AMR逐渐受到大家的关注，其他非英语语言的AMR语料库也得到了丰富，Li et al. (2016)将AMR推广到中文，称为中文抽象语义表示(Chinese Abstract Meaning Representation, CAMR)。近些年，荀恩东(2023)提出意合图理论，它是以事件为中心的语义表征图，为单根有向图，图中的节点对应承载事件、实体、属性的单元，边为有向边，表示单元间的语义关系，意合图力求能够对句子、段落、篇章等不同层级的语言单元作一贯式表示(郭梦溪et al., 2024)。

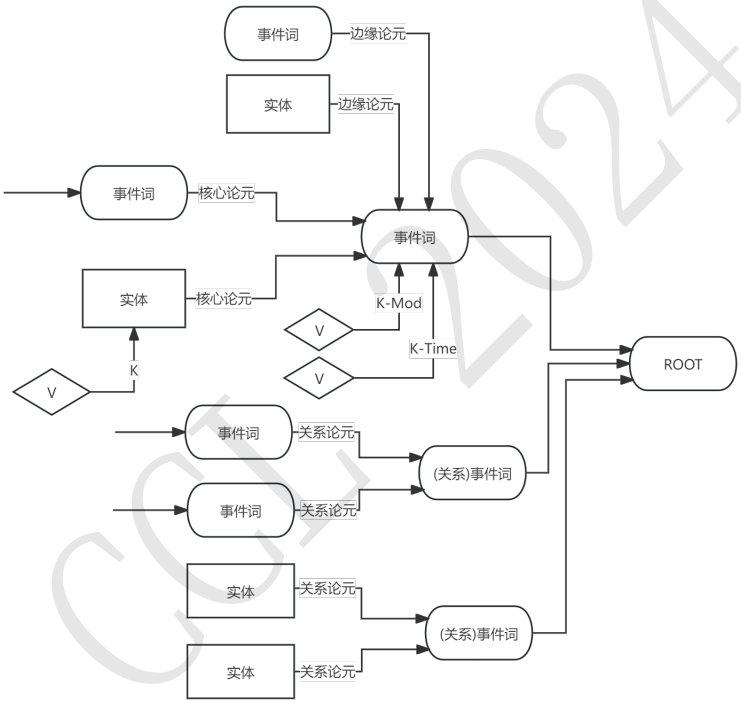

图 1: 意合图抽象表示

语义表示方法的发展也推动了语义解析技术的进步。目前，针对AMR语义解析的主流方法主要分为三类：基于图的方法(Graph-based)(Flanigan et al., 2014; Zhang et al., 2019)、基于转移的方法(Transition-based) (Wang et al., 2015; Ballesteros and Al-Onaizan, 2017; Astudillo et al., 2020)和基于序列到序列的方法(Seq2Seq-based)(Barzdins and Gosko, 2016; Peng et al., 2017; Noord and Bos, 2017; Konstas et al., 2017; Xu et al., 2020)。随着大模型时代的到来，一些研究将大模型技术引入语义解析任务中。Yang et al.(2023)探索了大型语言模型是否能够借鉴Seq2Seq建模方式在复杂结构化预测任务上得以运用。高文炀et al.(2023)选择微调Baichuan-7B模型来以端到端的形式从文本直接生成序列化的CAMR。

受此启发，我们选择了Llama3的中文微调模型Llama3-Chinese-8B-Instruct作为基座模型，在此基础上对其进行了LoRA (Hu et al., 2021)微调来进行中文意合图的语义解析。由于微调样本的不同设计会显著影响大模型微调后的效果，我们进行了样本设计工程(Sample Design Engineering, SDE)(Guo et al., 2024)设计。此外，在实验过程中，经过统计发现，大模型对不同语义标签解析难度不同，为此，我们设置了不同的模型训练策略，并对解析结果进行了组合分析。

2 方法

我们首先对中文意合图数据集进行了分析统计，根据统计规律以及任务性质进行了SDE设计。在设计好的样本集上微调Llama3-Chinese-8B-Instruct模型生成结构化的意合图表征三元组。为了提高大模型输出结果的质量和可靠性，我们设计了一些规则进行后处理。

2.1 SDE设计

根据简单的任务分析与统计，我们发现中文意合图数据集只有3000条可用的训练数据，共出现了63个语义标签，且除了要挖掘句子中词语之间的语义关系外，还需要对隐式事件词进行补充，所以我们认为中文意合图语义解析任务是一个较为复杂的下游任务。而细致地考虑大模型微调样本的设计，可以使用更少的样本训练出在下游任务上表现更好的模型(Guo et al., 2024)。因此，我们对评测提供的中文意合图数据集进行了简单分析，使用了SDE技巧来完成微调样本的设计。

图 2: 输入样本格式

最终，我们的大模型输入样本如图2所示。整个输入样本主要包含了四部分内容，分别为：上下文、指令、输入数据和输出指示。

- 上下文：对意合图的概念做了简单解释，帮助大模型更好的理解中文意合图的表示方法；
- 指令：概述模型需要执行意合图语义解析任务；
- 输入数据：告知模型需要进行语义解析的句子，句子的输入形式为json格式，“sent”字段中为按照特定方式分好词的句子；
- 输出指示：详细说明模型输出的格式，模型最后需要输出结构化的json格式，并向模型解释每个字段的意义和值域选项。

在输入样本格式的设计过程中，我们采用了Guo et al.(2024)工作中经过实验验证的SDE设计技巧。首先，我们将上下文、指令以及输出指示放置于输入的任务文本之前，这样更有助于提升模型的任务理解能力；其次，因为输出设计越格式化，格式输出错误的几率就越低，并且评测任务最终要求提交的是json格式，所以我们选择最结构化的输出格式json；最后，我们按照在训练集中出现的频次，从大到小排列语义标签，以提高模型对出现次数多的标签的关注程度。

2.2 大模型微调

Llama3-8B是一个基于仅解码器Transformer架构的多语言模型，拥有近80亿参数，在超过15万亿个标记(tokens)的公开数据上预训练，支持8192的上下文长度。通过引入分组查询注意力机制(Group Query Attention, GQA)(Ainslie et al., 2023)、扩大模型规模、更新分词器、增加词表大小和使用更为庞大的训练数据集，Llama3-8B展现出了强大的语言理解和生成能力。Llama3-Chinese-8B-Instruct1[1]是以Llama3-8B为基座模型的中文指令微调版本。我们使用LoRA(Hu et al., 2021)对Llama3-Chinese-8B-Instruct模型进行了监督微调，微调使用的参数设置如表1所示。

参数	参数值	参数含义
截断长度	4096	输入序列分词后的最大长度
学习率	5e-5	AdmaW优化器的初始学习率
训练轮数	6.0	需要执行训练总轮数
最大样本数	3000	每个数据集最多使用的样本数
计算类型	Fp16	训练使用的混合精度类型
批处理大小	1	批处理的样本数量

表 1: 微调参数设置

在微调过程中，我们以最小化交叉熵损失函数为优化目标：

$$J(\theta) = -\sum_{z_i \in V} z_i log(P(\hat{z}_i)) \tag{1}$$

$$P(\hat{z}_i) = \frac{e^{\hat{z}_i}}{\sum_{z_i \in V} e^{\hat{z}_i}} \tag{2}$$

其中，V是词汇表的大小，z_i为真实分布中第i个词的值（对于one-hot编码，目标词的z_i为1，其余为0），$\hat{z}_i$为模型预测该词的概率。

[1]https://www.modelscope.cn/FlagAlpha/Llama3-Chinese-8B-Instruct.git

模型	F1值
chatglm3-6b[1]	0.6572
Qwen1.5-7B[2]	0.6595
Qwen2-7B[3]	**0.6968**
Chinese-Falcon-7B[4]	0.4189
Llama3-Chinese-8B-Instruct	0.6251
Yi-1.5-9B[5]	0.6739
依存模型	0.3103
GPT4交互式	0.3729

表 2: 微调后不同基座模型以及基线的预测结果

2.3 后处理

大语言模型是概率模型，只是基于模型输入预测输出，因此大模型的输出不可避免地会产生一定的幻觉。而且，由于模型输出上下文长度有限，会导致大模型生成的json格式不完整。为了缓解这些问题对最终结果的影响，我们采取了以下三种后处理方式：

- json格式补全：在生成过程中，由于上下文长度的限制，模型生成的结果可能会被提前截断。对于这些无法直接进行json解析的预测结果，我们采用正则表达式匹配找到最后一个完整的三元组，舍弃剩余部分，并补全json格式，以尽可能多地保留模型生成的结果。

- 节点编号对齐：由于大模型对数字不敏感，在预测节点编号时可能会产生错误。为解决这一问题，我们采用规则的方法进行了节点编号的对齐处理。具体而言，如果节点内容词是隐式事件词，我们按照编号的对应关系直接对齐；如果节点内容词出现在输入句子中，则需考虑两种情况：1）若内容词在输入句子中仅出现一次，则将节点编号与内容词在输入句子中的索引对齐；2）若内容词在输入句子中出现多次，则将内容词定位到与输入句子中最接近预测编号的位置，节点编号与该处的索引对齐。

- 节点内容词修正：在模型生成的结果中，可能会出现生成的节点内容词既不是隐式事件词，又不在输入句子中的情况。针对这一情况，我们首先判断生成的节点编号是否合法，即该编号是否大于0且小于输入句子列表的长度。若编号合法，我们将生成的内容词替换为在输入句子中以该节点编号为索引的词；若编号不合法，则直接删除该节点的三元组。

3 实验

3.1 全语义标签微调

按照图2中的样本格式作为模型输入，在表1所示的参数设置下对Llama3-Chinese-8B-Instruct进行了监督微调。模型使用评测任务中的所有有标签中文意合图数据集作为训练集，对全部63个语义标签进行预测。为了评估其性能，我们在相同实验条件下微调了一系列同等规模的开源模型，实验结果如表2所示。由实验结果可知，微调后的Llama3-Chinese-8B-Instruct模型在该任务评测集上的F1值明显优于依存模型和GPT4交互式基线模型，且展现出了比Chinese-Falcon-7B模型更好的性能。然而，与其他中文模型相比，其性能仍存在不足。由此可见，在解决中文意合图语义解析任务方面，相较于使用中文语料指令微调的英文大模型，中文大模型展现出更高的适应性和优越性。

[1]https://www.modelscope.cn/ZhipuAI/chatglm3-6b.git
[2]https://www.modelscope.cn/qwen/Qwen1.5-7B.git
[3]https://www.modelscope.cn/qwen/Qwen2-7B.git
[4]https://huggingface.co/Linly-AI/Chinese-Falcon-7B
[5]https://www.modelscope.cn/01ai/Yi-1.5-9B.git

此外，为了评估大模型对不同语义标签的预测能力，我们在微调后的Llama3-Chinese-8B-Instruct模型上对3000条训练集数据的进行了推理预测。定义语义标签l的错误率err_l，该错误率衡量的是语义标签l的预测准确性，具体公式如下：

$$err_l = \frac{\sum_{i=1}^{3000} |\{(word1, word2, relVal) \in P_i \cap T_i : relVal = l\}|}{\sum_{i=1}^{3000} |\{(word1, word2, relVal) \in T_i : relVal = l\}|} \tag{3}$$

其中，P_i表示第i个样例的预测结果中含有语义标签l的三元组集合，T_i表示第i个样例的真实结果中含有语义标签l的三元组集合。

各语义标签的错误率如图3所示。根据实验结果，我们发现，在全语义标签微调下，大模型对不同语义标签的预测能力存在差异。因此，基于错误率是否小于0.5，我们将语义标签分为易预测和难预测两类，并进行了进一步实验。

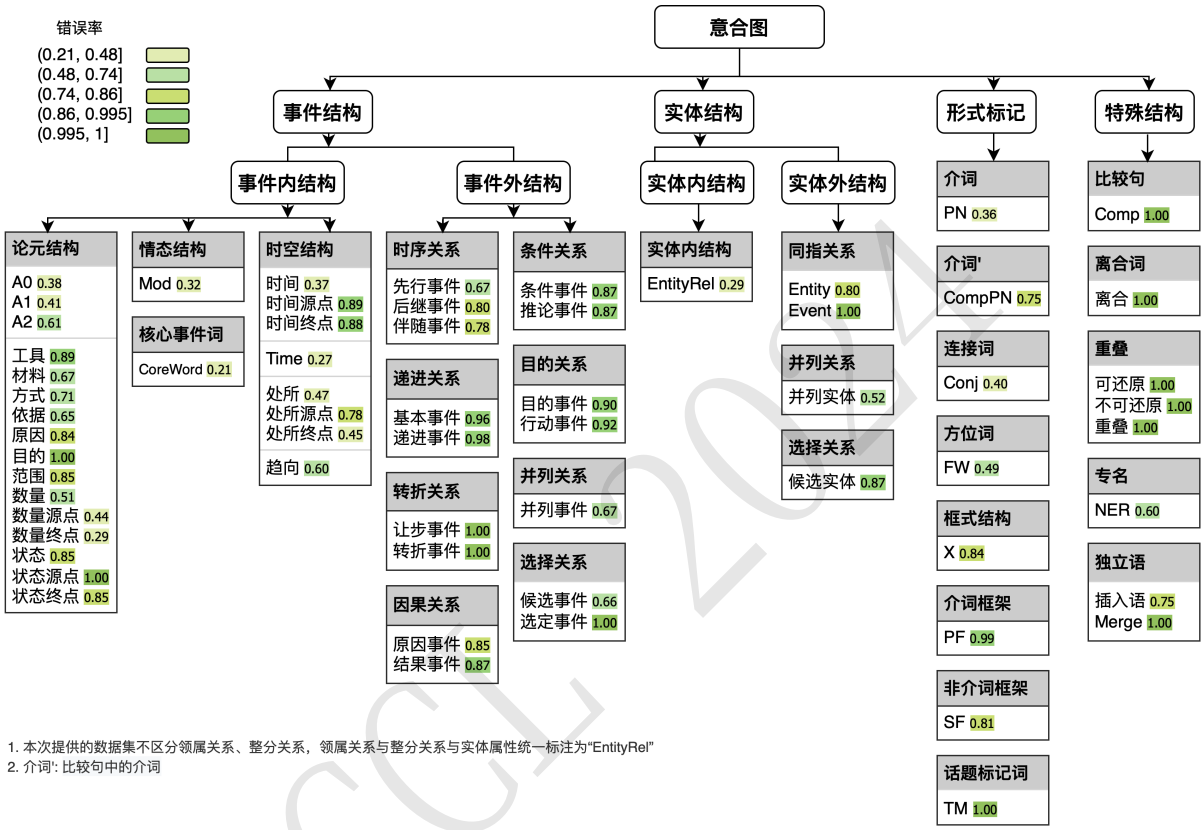

图 3: 全语义标签微调模型在训练集上各语义标签的错误率

3.2 语义标签分类微调

在本节中，我们针对易预测(l_easy)与难预测(l_hard)两种语义标签类别，将整个数据集划分为简单集与困难集两部分，简单数据集中仅包含标签类别为" l_easy"的三元组，而困难数据集则仅包含标签类别为"l_hard"的三元组。我们分别在这两部分数据集进行了Llama3-Chinese-8B-Instruct模型的监督微调。为了得到更优的语义解析结果，我们对不同微调模型的预测结果按照语义标签类别进行了组合，实验结果如表3所示。

由实验结果可知，难预测的语义标签预测结果的F1值非常低，进一步探究其原因，我们发现一些语义标签（如"Merge" "状态源点" "可还原" "目的" "离合" "重叠" "不宜还原" "TM" "选定事件"）在训练集中只有不到25条数据，少者甚至只有1条。这些语义标签由于训练数据的缺乏，导致训练不充分，从而对模型的理解和预测造成困难。此外，一些语义标签的训练数据并不少（如"行动事件" "目的事件" "结果事件"等），但其预测准确率依然很低。通过对这些语义标签的更进一步分析，我们发现，标注为这些语义标签的两个词中往往有一个词为"隐性事件词"。如在下面例句中，"出去"和"目的关系"标注为"行动事件"，其中"目的关系"则是"隐性事

模型	语义标签	F1值
Total	L_total	0.625
	L_easy	0.606
	L_hard	0.109
Easy	L_easy	0.629
Hard	L_hard	0.074
Easy	L_easy	0.625
Hard	L_hard	
Total	L_hard	**0.646**
Easy	L_easy	
Total	L_easy	0.604
Hard	L_hard	

表 3: 不同微调模型的预测结果。其中，Total为全语义标签微调的模型，Easy为在简单数据集上训练的模型，Hard为在困难数据集上训练的模型；L_total为全部语义标签的集合，L_easy为属于"*l_easy*"类别的语义标签集合，L_hard为属于"*l_hard*"类别的语义标签集合。

件词"，若要成功预测生成该类语义标签下的三元组，则须先将"隐性事件词"识别并抽取出来。因此，我们认为这些语义标签对大模型而言较难学习和掌握。

例句：

['老王','退休','以后','觉得','生活','没有','意思',',',',','所以','妻子','让','他','出去','找','点儿','事','干','。']

{'word1': 'word': '出去', 'idx': 12,'word2': 'word': '目的关系', 'idx': -10,'relVal': '行动事件'}

我们还发现，使用简单数据集进行微调有助于提高易预测语义标签的预测性能，然而，使用困难数据集微调反而降低了难预测语义标签预测的F1值，我们推测，这可能是因为模型在预测其他语义标签时学到了意合图语义解析的一般规律，这对预测难预测语义标签是有益的。

4 结语

在本次评测任务中，我们利用SDE进行了输入样本格式设计，并在此基础上对开源大模型进行了微调，以解决中文意合图语义解析任务。实验结果表明，使用中文语料指令微调的英文大模型解决该任务的能力逊于中文大模型，并且大模型对不同语义标签的预测能力存在显著差异。最终，我们的参赛系统在评测集上F1值达到了0.6461，在该任务中取得了第三名的成绩。然而，该任务仍存在一些需要进一步研究的问题。未来，我们将更加深入地分析大模型在不同标签上预测能力不同的原因，并将继续探索如何提高大模型在难预测语义标签上的预测性能。

参考文献

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. *In Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Bin Li, Yuan Wen, Weiguang Qu, Lijun Bu, and Nianwen Xue. 2016. Annotating the little prince with Chinese AMRs. *In Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 7–15, Berlin, Germany. Association for Computational Linguistics.

荀恩东. 2023. 自然语言结构计算：意合图理论与技术. 人民邮电出版社.

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. 2014. A discriminative graph-based parser for the Abstract Meaning Representation. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* , volume 1, pages 1426–1436, Baltimore, Maryland. Association for Computational Linguistics.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019. AMR Parsing as Sequence-to-Graph Transduction. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, volume 1.

Chuan Wang, Nianwen Xue, and Sameer Pradhan. 2015. Boosting Transition-based AMR Parsing with Refined Actions and Auxiliary Analyzers. *In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–375, Denver, Colorado. Association for Computational Linguistics.

Miguel Ballesteros, and Yaser Al-Onaizan. 2017. AMR Parsing using Stack-LSTMs. *ArXiv, abs/1707.07755*.

Ramo n Fernandez Astudillo, Miguel Ballesteros, Tahira Naseem, Austin Blodget, and Radu Florian. 2020. Transition-based Parsing with Stack-Transformers. *ArXiv, abs/2010.10669*.

Guntis Barzdins, and Didzis Gosko. 2016. RIGA at SemEval-2016 Task 8:Impact of smatch extensions and character-level neural translation on AMR parsing accuracy. *In Proceedings of International Workshop on Semantic Evaluations (SemEval)*, pages 1143-1147, San Diego, USA. Association for Computer Linguistics.

Xiaochang Peng, Chuan Wang, Daniel Gildea, and Nianwen Xue. 2017. Addressing the data sparsity issue in neural AMR parsing. *In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 366–375, Valencia, Spain. Association for Computational Linguistics.

Rik van Noord, and Johan Bos. 2017. Neural Semantic Parsing by Character-based Translation: Experiments with Abstract Meaning Representations. *ArXiv, abs/1705.09980*.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural AMR: Sequence-to-sequence models for parsing and generation. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 146-157.

Dongqin Xu, Junhui Li, Muhua Zhu, Min Zhang, and Guodong Zhou. 2020. Improving AMR parsing with sequence-to-sequence pre-training. *In Proceedings of the EMNLP*, pages 2501-2511.

Yifei Yang, Ziming Cheng , and Hai Zhao. 2023. CCL23-Eval任务2系统报告：基于大型语语言模型的中文抽象语义表示解析.

高文炀, 白雪峰, and 张岳. 2023. CCL23-Eval任务2系统报告：WestlakeNLP，基于生成式大语言模型的中文抽象语义表示解析.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv, abs/2106.09685*.

Biyang Guo, He Wang, Wenyilin Xiao, Hong Chen, Zhuxin Lee, Songqiao Han, and Hailiang Huang. 2024. Sample Design Engineering: An Empirical Study of What Makes Good Downstream Fine-Tuning Samples for LLMs. *ArXiv, abs/2404.13033*.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. *ArXiv, abs/2305.13245*.

郭梦溪，荀恩东，李梦，and 饶高琦. 2024. 意合图：中文多层次语义表示方法. 第二十三届中国计算语言学大会