Aligning Instruction Tasks Unlocks Large Language Models as Zero-Shot Relation Extractors

Kai Zhang **Bernal Jiménez Gutiérrez** **Yu Su**

The Ohio State University

{zhang.13253, jimenezgutierrez.1, su.809}@osu.edu

Abstract

*警告：该PDF由GPT-Academic开源项目调用大语言模型+Latex翻译插件一键生成，版权归原文作者所有。翻译内容可靠性无保障，请仔细鉴别并以原文为准。项目Github地址 https://github.com/binary-husky/gpt_academic/。当前大语言模型: gpt-4，当前语言模型温度设定: 1。为了防止大语言模型的意外谬误产生扩散影响，禁止移除或修改此警告。

近期的研究表明，对大型语言模型（LLMs）进行大规模指令跟踪数据集的微调，可以显著提高它们在各种NLP任务上的性能，特别是在零样本环境中。然而，即使是先进的指令调谐过的LLMs，其在关系抽取（RE）这一基础信息抽取任务上的表现仍无法超越小型语言模型。我们假设这是因为在指令调谐数据集中，RE的发生率很低，占所有任务总数的1%不到(Wang et al., 2022)。为解决这个问题，我们提出了QA4RE，这是一个将RE与问题回答（QA）进行对齐的框架，QA在指令调谐数据集中是一项主要任务。对两系列指令调谐过的LLMs（总共六个LLMs）在四个数据集上进行的全面零样本RE实验表明，我们的QA4RE框架一直都能提升LLMs的性能，这强烈证明了我们的假设，并使得LLMs在零样本基线上获得了大幅度的优势。此外，我们提供了详尽的实验和讨论，显示了我们QA4RE框架的健壮性、少样本效果和强大的传递性。这项工作揭示了通过将这些任务与更常见的指令调谐任务（如QA）进行对齐，从而使LLMs适应富有挑战性和代表性不足的任务的一种有前途的方法。[1]

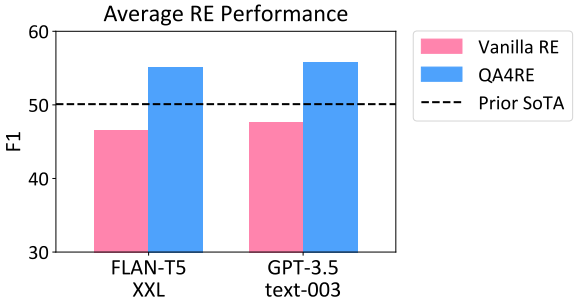

图 1: 主要发现：在标准（原始）的关系抽取（RE）框架下，强烈指令调通的大型语言模型（LLMs）在性能上不如之前的零次学习关系抽取方法。我们的QA4RE框架让两套指令调通的大型语言模型（分别是FLAN-T5和GPT-3.5）在4个关系抽取数据集上大幅度超越了之前的最先进水平。结果是对4个关系抽取数据集的平均值。在展示GPT-3.5模型时，我们省略了“davinci”这个词以简化。

1 Introduction

大型语言模型 (LLMs) (Brown et al., 2020; Chowdhery et al., 2022; Zhang et al., 2022) 在许多自然语言处理 (NLP) 任务中表现出令人印象深刻的性能。采用上下文学习范式，而无需任何参数更新，LLMs 能够达到与经过数千示例微调的小型语言模型 (LMs) (Liu et al., 2022; Min et al., 2022a; Liang et al., 2022) （我们将参数少于10亿的LMs视为小型模型）相媲美的性能。最近，通过在包含数千下游任务的数据集上微调LLMs，并将这些任务转化为指令执行格式（即指令调优），已经被证明可以在各方面大幅度提升LLMs，特别是在零样本设置下 (Iyer et al., 2022; Ouyang et al., 2022; Chung et al., 2022)。

我们研究了LLMs在识别句子中实体关系，即关系抽取 (RE)，一个在信息抽取中的基础

[1]代码和数据可在 https://github.com/OSU-NLP-Group/QA4RE 获取。

任务中的能力。最近的工作 (Jimenez Gutierrez et al., 2022) 发现在生物医学领域的RE任务上，LLMs的性能不如经过微调的小型LMs。我们在图 1 中对一般领域的RE进行的结果显示，即使是两种最先进的经过指令调优的LLMs，FLAN-T5 XXL (Chung et al., 2022) 和 text-davinci-003 (Ouyang et al., 2022)，也无法超过基于小型LMs的最先进的（SoTA）零样本RE方法 (Sainz et al., 2021)。

我们假设指令调优LLMs在RE能力有限可能是由于指令调优数据集中RE任务出现的频率低 (Ouyang et al., 2022; Sanh et al., 2022; Chung et al., 2022; Wang et al., 2022) （RE领域的任务在最大的现有指令数据集 (Wang et al., 2022)中的比例<0.5%；具体详情请查阅附录 A）。为了解决低发生率问题，我们提出了QA4RE框架，它将RE与多项选择问题回答 (QA) 对齐，后者在大多数指令调优数据集中出现的频率要高得多——在Wang et al. (2022) 和 Ouyang et al. (2022) 的任务中占所有任务的12-15%。具体来说，通过将输入句子视为问题，可能的关系类型视为多项选择选项（见图 2），LLMs能够通过选择代表正确关系类型的选项来进行RE。

在四个实际关系抽取数据集和两种不同系列的六个指令调优模型（OpenAI GPT-3.5 和 FLAN-T5 (Chung et al., 2022)）上的全面评估显示，QA4RE相比标准的RE公式带来了显著的提升，验证了其有效性和我们关于RE低发生率的假设。更具体地说，我们的框架使得text-davinci-003 和 FLAN-T5-XXLarge的F1平均分数分别获得了8.2%和8.6%的绝对改进。首次有一种LLM能够在零样本设置中显著超过了之前基于小型 LM 的SoTA。深入分析进一步展示了QA4RE的鲁棒性和少样本有效性。更重要的是，我们的框架已被证明可以在各种大小（从80M到175B）的指令调优模型中有效地转移应用。我们的贡献总结如下：

(1) 我们系统地研究了在四个实际关系抽取数据集上的指令调优LLMs，并注意到他们在RE上的限制性表现可能源于指令调优数据集中RE任务的低发生率。

(2) 我们将RE重构为多项选择QA，以尝试合理地利用QA在指令调优数据集中的更高频率，并在最近的六个指令调优LLMs上取得了显著的提升，首次显著超过了之前基于小型 LM 的SoTA零样本RE方法。

(3) 此外，我们展示了我们的QA4RE方法对于多样化提示设计的鲁棒性及其在少样本设置中的良好结果。**(4)** 最后，我们展示了QA4RE框架的有效性是可转移和一致的，适用于从80M到175B不同大小的指令调优模型。我们的研究揭示了将不常见和具有挑战性的任务与常见的指令调优任务对齐的潜力，并可以指导他人探索这一方向。

2 Related Work

指令调优。 大型语言模型最初是通过在大规模范围内利用自监督的下一个标记预测来获得令人印象深刻的零射和少数射性能。最近，如(Ouyang et al., 2022; Iyer et al., 2022; Wei et al., 2022a; Chung et al., 2022; Sanh et al., 2022)所示，针对大量下游任务进行监督细调已被展示出能提高LLM精度、健壮性、公平性以及对未见任务的泛化。一些策略已被开发出来以便将LLMs与人类指示对齐，包括来自人类反馈的强化学习（RLHF）(Ouyang et al., 2022)以及更常规的语言建模目标，将其用于在将一系列任务改写为遵循指示任务上对LLMs进行精调(Iyer et al., 2022; Wei et al., 2022a; Chung et al., 2022; Sanh et al., 2022)。

引出LLM能力。 LLM预训练的高成本和日益私有的性质使得最终确定不同预训练技术如何导致LLM不同能力的问题相当具有挑战性。在预训练中涉及的许多因素，如简单的自我监督扩展、代码或多语言文本预训练(Chowdhery et al., 2022; Chen et al., 2021; Chung et al., 2022)以及上文提到的不同版本的指令调优(Ouyang et al., 2022; Iyer et al., 2022; Wei et al., 2022a; Chung et al., 2022)，都可以通过多种方式相互

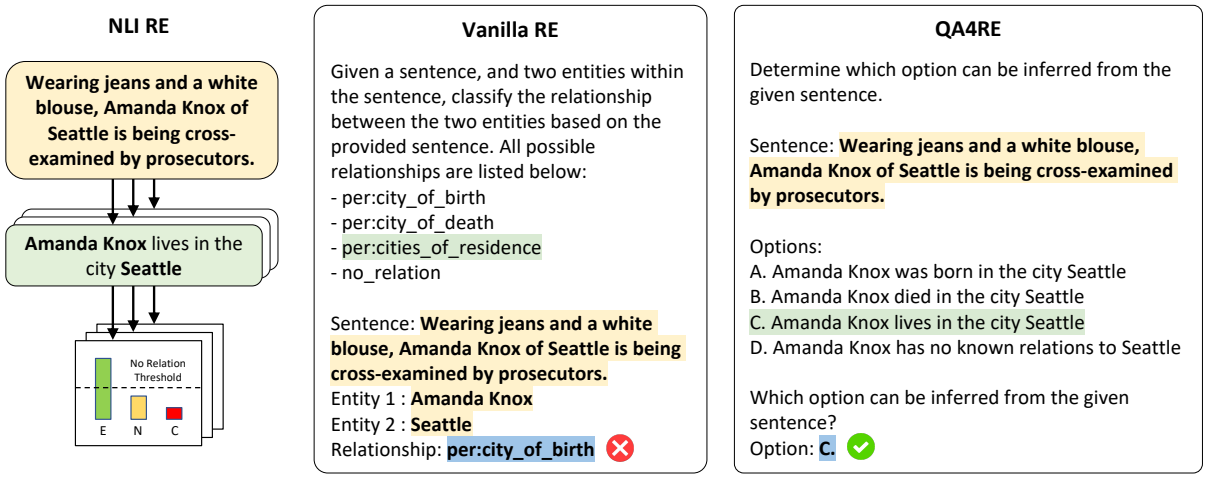

图 2: 此图展示了SoTA NLI零射击框架的概要，其中每个句子都必须与每个关系模板（左图）进行比较，Jimenez Gutierrez et al. (2022) 中进行RE的GPT-3的初级形式化（中图），以及我们的多选题QA设定，其中每个关系被转化为一个模板，然后预期GPT-3只预测一个单字母（右图）。

交互，释放出LLMs所表现出来的能力。然而，Fu and Khot (2022)假设在预训练中使用代码似乎可以提高LM的推理能力，这得到证实的是由部分基于PaLM (Chowdhery et al., 2022)，如code-davinci-002 (Chen et al., 2021)，以及text-davinci-002/003 (Ouyang et al., 2022)这类语言模型的Chain-of-Thought prompts能力的提高，与像text-davinci-001和OPT-175B (Zhang et al., 2022)这样的仅文本模型相比。此外，在大量任务上进行指导调优已被显示能够改善对未见任务的泛化，减少对少射例子的需要，并提高许多语言任务的准确性和健壮性(Ouyang et al., 2022; Iyer et al., 2022; Chung et al., 2022)。

低资源关系抽取。 几种对标准RE的改写使得小型LMs能够在零射和少数射设置中达到相当强的性能。 Sainz et al. (2021)利用经过自然语言推理（NLI）数据集细调的小型LMs来通过选择大多数由测试句子由来的实体填充关系模板来进行零射RE。 Lu et al. (2022)将RE定义为总结任务，并借助生成模型来总结句子中目标实体之间的关系。其他低资源RE方法通过使用逻辑规则从子-提示创建复杂提示(Han et al., 2022)和使用可学习的虚拟标记注入有关实体类型的知识来增强提示调整(Chen et al., 2022)。我们当前的工作使用了这些研究中设计的几种关系模板。

LLMs用于关系抽取。 在探索LLMs的RE能力方面，大多数先前的工作都侧重于研究生物医学RE。 Jimenez Gutierrez et al. (2022)报告称LLMs在全面的生物医学RE数据集的少数射设置中的表现低于经过标准小型LMs细调的理想表现，并且显示出对none-of-the-above(NoTA)关系类别的处理不彻底是其中的一大误解。进一步来说，尽管Super Natural Instruction (Wang et al., 2022)包含了一些类似RE的任务，但这些任务占据了数据集的大约0.5%且都没有被选择用于模型评估。

3 Methodology

在本节中，我们正式定义了关系抽取问题，并详细描述了我们对该问题的多项选择问答方法。

3.1 Problem Statement

关系抽取 (RE) 的目标是基于特定句子抽取两个给定实体之间的关系。更具体地说，一个关系实例包含一个句子 S，以及 S 中的一个头实体 E_h 和一个尾实体 E_t。给定一个关系实例 (S, E_h, E_t)，模型需要从一组预定义的关系类型中识别 S 中表达的 E_h 与 E_t 之间的关系。

3.2 Relation Templates

最近的低资源RE方法 (Sainz et al., 2021; Lu et al., 2022; Han et al., 2022) 使用关系蕴含模板作为标签口语化（例如，“per:city_of_birth” -> “{E_h} 在城市{E_t}出生”）。如图 2（左）所示，当前的SoTA方法针对低资源RE (Sainz et al., 2021) 使用手动构建的关系模板来将RE任务重新构 formulate成自然语言推理（NLI）任务。

为确保公正的比较，我们使用之前研究中开发的相同模板 (Sainz et al., 2021; Lu et al., 2022) 在我们的QA4RE框架中生成答案选项。此外，在 Sec. 6.2 中，我们讨论了直接将NLI形式应用于LLMs中的RE的可能性。

3.3 QA4RE Framework

如图 2（右）所示，我们将关系抽取任务重新构 formulated 为多选 QA 问题。通过将给定的头和尾部关系实体（E_h 和 E_t）集成到关系模板中，并将它们用作多选项，LLMs能够利用广泛的 QA 指导进行微调，这大大提高了最近的模型。此外，我们的方法允许 LLM 只生成答案索引，而不是像在之前的工作中 verbalize 关系 (Jimenez Gutierrez et al., 2022)，如图 2（居中）所示。

类型约束答案构造。 为了将 RE 转化为一道多选题，对于给定的关系实例 (S, E_h, E_t)，我们使用句子 S 作为标准 QA 的上下文，并创建由预定义模板填充的 E_h 和 E_t 实体组成的选项。为了公平地比较以前的工作，我们应用类型约束（如果适用）来消除与头部和尾部实体的实体类型不兼容的关系类型的选项。例如，如果 E_h 的类型是 PERSON，那么关系 "org:country_of_headquarters" 将被认为是无效的，因为一个人没有总部。

4 Experiment Setup

4.1 Datasets

我们在四个RE数据集上评估了我们的方法： (1) TACRED (Zhang et al., 2017)， (2) RE-TACRED (Stoica et al., 2021)， (3) TACREV (Alt et al., 2020)，以及 (4) SemEval 2010任务8（简称SemEval） (Hendrickx et al., 2010)。跟随之前的工作 (Sainz et al., 2021; Lu et al., 2022; Han et al., 2022; Chen et al., 2022)，我们报告了除了“无以上关系”以外的微平均F1。为了控制OpenAI API的费用，我们从各数据集的测试划分中随机抽取了1,000个样例作为我们的测试集。

4.2 Baselines

零样本学习. 对于小型的基于LM的模型，我们评估了两种低资源SoTA RE基线: (1) 如图 2 (左)所示，NLI (Sainz et al., 2021) 将RE做为一种自然语言推理任务进行重新构思，并利用了在MNLI数据集 (Williams et al., 2018)上进行微调的几个LM：BART-Large (Lewis et al., 2020)，RoBERTa-Large (Liu et al., 2019)，和DeBERTa-XLarge (He et al., 2021)。这种方法在零样本和少样本RE上都保持了最好的效果。 (2) 此外，SuRE (Lu et al., 2022) 将RE视为一种摘要任务，并利用了生成型LMs，如BART-Large (Lewis et al., 2020)和PEGASUS-Large (Zhang et al., 2020)，在少样本和全监督设置中获得了竞争性的结果。

对于NLI方法 (Sainz et al., 2021)，我们在TACRED 和 TACREV 上报告了使用它们自己的模板的性能。由于该方法没有为RETACRED 和 SemEval 设立模板，我们在这两个数据集上使用了后续的工作SuRE (Lu et al., 2022)的模板。所有零样本方法，包括那些基于LLM的方法，都应用实体类型限制来减小关系标签空间。由于SemEval没有提供实体类型，上述方法在每个实例中使用所有可能的关系作为标签空间。

少样本学习. 尽管我们的主要实验关注的是零样本的RE，但我们还进一步探索了我们的方法通过比较其在TACRED数据集上的少样本性能与几种竞争性的小型基于LM的方法的能力。

NLI基线可以很容易地扩展到少样本设置。[2] 此外，我们添加了(1) 标准的Fine-Tuning (Jimenez Gutierrez et al., 2022), (2) 使用逻辑规则的提示调整的PTR (Han et al., 2022)，以及(3) 通过学习虚拟令牌使用实体类型知识的KnowPrompt (Chen et al., 2022)，所有这些都是用RoBERTa-Large (Liu et al., 2019)进行初始化的。有关超参数详情，请参阅附录 B.1。

4.3 QA4RE Implementation Details

我们的QA4RE框架利用了先前工作(Sainz et al., 2021; Lu et al., 2022)开发的相同模板和类型约束。具体来说，我们在所有4个数据集上使用SuRE (Lu et al., 2022)模板进行我们的QA4RE方法，因为NLI (Sainz et al., 2021)模板只是为TACRED设计的。对于提示工程，我们在初步实验中探索了用于普通RE和QA4RE的提示格式和任务指示，使用$text-davinci-002$进行在一个包含250个示例的TACRED开发集的子集上的实验。然后，我们使用同样的任务指示和提示格式用于所有四个数据集和LLMs。关于提示格式和关系口词化模板的详细内容，请参见附录B.2和B.3。

为了系统地比较我们的QA4RE框架与普通的RE形式，我们在两系列的LLMs上对他们进行评估，总共得出了七个模型。在GPT-3.5系列的LLMs中，对于可以通过文本完成API(code-davinci-002，text-davinci-002和text-davinci-003)访问的LLMs，我们遵循以前的工作(Jimenez Gutierrez et al., 2022)，并使用logit偏差选项来限制令牌生成为普通RE和QA4RE的选项索引。由于在聊天完成API(gpt-3.5-turbo-0301)中LLMs的可控选项较少，我们只将温度设为0并使用默认的系统提示。

我们还检查了开源的经过一系列任务(Sanh et al., 2022; Wei et al., 2022a; Wang et al., 2022)训练的FLAN-T5系列LLMs (Chung et al., 2022)。在训练中使用的1,836个任务中，只有不到0.5% 的任务类似于RE，使得FLAN-T5系列的LLMs成为验证我们假设的理想模型。具体来说，我们使用XLarge(3B)和XXLarge(11B)模型，并采用与GPT-3.5系列LLMs相同的提示和贪婪解码策略来确保公平的比较。

5 Results

5.1 Zero-Shot Results

我们在四个关系抽取数据集上的主要实验结果可以在Tab. 1中找到。我们从结果中有以下观察：

(1) 通过将RE重塑为QA，我们的框架在所有LLM和大部分数据集上提升了既有的RE形式，使它们成为更强的零射击关系抽取器。特别的，text-davinci-003和FLAN-T5 XL和XXL能够以较大的优势超过先前的SoTA，$\text{NLI}_{\text{DeBERTa}}$。值得注意的一点是，QA4RE在每个系列中表现最好的LLM（text-davinci-003和FLAN-T5 XXL）上带来了最大的提升，这表明更强的LLM可能会更多地受益于我们的框架。

(2) Tab. 1中的两个FLAN-T5 LLM显著地从我们的QA4RE框架中受益。此外，在其他的FLAN-T5模型以及完整的测试集中也可以观察到持续并且显著的提升，正如在Sec. 6.3和附录C中所讨论的。考虑到关系抽取任务占用训练FLAN-T5模型的教程任务不到0.5%，这些发现强烈支持我们的假设，即将代表性不足的任务与更常见的教程调试任务（如QA）对齐，可以解锁LLM处理低频率任务的能力。

(3) SemEval数据集对所有的基线都提出了显著的挑战，尤其是对于缺乏类型约束的SuRE (Lu et al., 2022)。对于这样一个大的搜索空间，没有经过精调的生成LMs倾向于将所有的例子总

[2] SuRE也可以扩展到少样本设置,但我们无法使用提供的代码复制他们的结果。

Methods		TACRED			RETACRED			TACREV			SemEval			Avg.
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	F1
Baselines														
NLI_{BART}		42.6	65.0	51.4	59.5	34.9	44.0	44.0	74.6	55.3	21.6	23.7	22.6	43.3
$\text{NLI}_{\text{RoBERTa}}$		37.1	76.9	50.1	52.3	67.0	58.7	37.1	83.6	51.4	17.6	20.9	19.1	44.8
$\text{NLI}_{\text{DeBERTa}}$		42.9	76.9	55.1	71.7	58.3	64.3	43.3	84.6	57.2	22.0	25.7	23.7	50.1
$\text{SuRE}_{\text{BART}}$		13.1	45.7	20.4	17.9	34.6	23.6	14.1	52.3	22.2	0.0	0.0	0.0	16.5
$\text{SuRE}_{\text{PEGASUS}}$		13.8	51.7	21.8	16.6	34.6	22.4	13.5	54.1	21.6	0.0	0.0	0.0	16.4
GPT-3.5 Series														
ChatGPT	Vanilla	32.1	74.8	44.9	45.4	61.3	52.1	30.3	79.6	43.9	18.2	20.8	19.4	40.1
	QA4RE	32.8	68.0	44.2 (−0.7)	48.3	76.8	59.3 (+7.2)	34.7	79.1	48.2 (+4.3)	29.9	35.2	32.3 (+12.9)	46.0 (+5.9)
code-002	Vanilla	27.2	70.1	39.2	42.7	70.4	53.1	27.5	77.7	40.6	27.2	25.6	26.4	39.8
	QA4RE	37.7	65.4	47.8 (+8.6)	48.0	74.0	58.2 (+5.1)	31.7	65.5	42.7 (+2.1)	25.2	29.2	27.0 (+0.6)	43.9 (+4.1)
text-002	Vanilla	31.2	73.1	43.7	44.1	76.3	55.9	30.2	76.8	43.3	31.4	28.8	30.1	43.2
	QA4RE	35.6	68.4	46.8 (+3.1)	46.4	72.4	56.5 (+0.6)	35.7	76.8	48.8 (+5.4)	29.4	34.3	31.6 (+1.5)	45.9 (+2.7)
text-003	Vanilla	36.9	68.8	48.1	49.7	62.2	55.3	38.2	76.8	51.0	33.2	39.3	36.0	47.6
	QA4RE	47.7	78.6	**59.4** (+11.3)	56.2	67.2	61.2 (+5.9)	46.0	83.6	**59.4** (+8.4)	41.7	45.0	43.3 (+7.3)	**55.8** (+8.2)
FLAN-T5 Series														
XLarge	Vanilla	51.6	49.1	50.3	54.3	40.3	46.3	56.0	59.1	57.5	35.6	29.8	32.4	46.6
	QA4RE	40.0	78.2	53.0 (+2.7)	57.1	79.7	66.5 (+20.2)	40.7	85.9	55.3 (−2.2)	45.1	40.1	42.5 (+10.1)	54.3 (+7.7)
XXLarge	Vanilla	52.1	47.9	49.9	56.6	54.0	55.2	52.6	50.9	51.7	29.6	28.8	29.2	46.5
	QA4RE	40.6	82.9	54.5 (+4.6)	56.6	82.9	**67.3** (+12.1)	39.6	86.4	54.3 (+2.6)	41.0	47.8	**44.1** (+14.9)	55.1 (+8.6)

表 1: 在四个RE数据集上的实验结果（%）。我们省略了GPT-3.5系列LLMs的名称中的'davinci'，ChatGPT指的是gpt-3.5-turbo-0301。我们将最佳结果标记为**粗体**，次佳结果为下划线，并且我们的QA4RE相对于原生RE的F1提升用绿色表示。

结为NoTA关系，导致了它的系统性失败。需要注意的是，没有类型约束，RE问题在我们的QA4RE框架中变成了一个有19个选择的问答任务。尽管如此，我们的方法仍然在LLM上显示出显著的改进，特别是对于text-davinci-003和FLAN-T5 XXL。

5.2 Robustness to Verbalization Templates

对于我们的实验，我们使用了之前的工作中手动编写的关系模板 (Sainz et al., 2021; Lu et al., 2022)。然而， Lu et al. (2022)指出，模型的性能可能会因模板设计的不同而有很大的差异。因此，为了研究模型对不同模板的鲁棒性，我们使用了四种不同的模板进行了全面的实验，这些模板在附录 B.3中有详细描述，这些实验覆盖了TACRED数据集上的所有零射击方法。表 2展示了在我们的主要实验中使用的所有方法上，这四种模板的比较结果，包括作为无模板参考的原始RE。

从表2中，我们可以观察到以下几点：

(1) 无论模板如何，我们的方法始终优于小型LM基线和普通的RE框架。值得注意的是，即使使用仅含有标签名信息而无专家知识构建的模板（TEMP3和TEMP4），我们的QA框架的表现仍优于普通的RE，表明我们的QA框架的有效性和一致性。

Methods		TEMP1	TEMP2	TEMP3	TEMP4
NLI_{BART}		51.4	49.7	4.4	42.0
$\text{NLI}_{\text{RoBERTa}}$		50.1	47.1	19.6	35.8
$\text{NLI}_{\text{DeBERTa}}$		55.0	49.4	17.1	36.6
$\text{SuRE}_{\text{BART}}$		19.9	20.4	2.1	10.1
$\text{SuRE}_{\text{PEGASUS}}$		20.5	21.8	6.2	19.3
text-003	Vanilla	48.1			
	QA4RE	**56.6**	**59.4**	**48.7**	**50.1**

表 2: 在TACRED上的F1评分，使用四个模板（%）。每个模板的最佳结果都以粗体标出。text-003指的是text-davinci-003。

(2) NLI和SuRE的性能在很大程度上依赖于模板。当使用精心制作的高质量模板（TEMP1和TEMP2）时，几种基于LM的NLI方法的性能优于使用普通RE的text-davinci-003。然而，当使用无需专家知识创建的模板（TEMP3和TEMP4）时，NLI和SuRE的性能显著下降。相比之下，QA4RE对于口头化模板的变化更为稳健，降低了试错开发的努力，同时也使其更容易转移到由于专家注释的高成本而可能限制获取质量模板的设置，如生物医药或金融领域。

5.3 None-of-the-Above Relation Evaluation

"none-of-the-above"（NoTA）关系 (Gao et al., 2019; Sabo et al., 2021; Jimenez Gutierrez et al., 2022)被定义为给定实体之间不存在任何感兴趣的关系的情况。 Jimenez Gutierrez et al. (2022)展示了LLMs在RE任务上早期较差的表现很大程度上可以归因于它们无法处理NoTA关系。为了评估零射击方法在NoTA关系上的效力，遵循之前的工作 (Fei and Liu, 2016; Shu et al., 2017; Sainz et al., 2021)，我们应用包含NoTA的宏F1指标，以及微观和宏观P vs. N (所有正面关系 vs. NoTA关系作为二元分类）的F1指标。

Methods		Macro F1	Micro P vs. N	Macro P vs. N
NLI_{BART}		49.8	75.9	71.1
$\text{NLI}_{\text{RoBERTa}}$		43.7	68.5	65.8
$\text{NLI}_{\text{DeBERTa}}$		55.0	75.6	72.3
$\text{SuRE}_{\text{BART}}$		15.5	35.2	35.0
$\text{SuRE}_{\text{PEGASUS}}$		14.9	32.4	31.5
text-003	Vanilla	45.3	72.8	69.5
	QA4RE	**58.9**	**78.4**	**74.8**

表 3: 包含 NoTA 的 42-class 宏 F1，以及宏和微 P vs N（所有积极关系 vs NoTA）在 TACRED（%）上的 F1。每个度量的最佳结果都被加粗表示。text-003 指的是 text-davinci-003。Ma 和 Mi 分别是宏观(macro)和微观(micro)的缩写。

从表 3中，我们观察到，通过我们的QA框架增强，text-davinci-003在包含NoTA的指标上取得了显著的改进，超过了基于小LM的NLI方法。这进一步证明了我们框架的有效性，即使在处理具有挑战性的NoTA关系上也不例外。值得注意的是，这些优越的结果是通过简单地在QA中添加一个实体填充的NoTA关系模板作为答案选项，而无需前一种方法 (Sainz et al., 2021; Lu et al., 2022)的额外阈值要求。这消除了对额外超参数搜索的需求，这种搜索对于低资源设置可能会很棘手。

5.4 Few-Shot Results

虽然零射程RE是我们的主要关注点，但我们也在少射程设置下评价我们的方法。结果如图Tab. 4所示。由于预算限制，我们将我们的案例研究限制在4射程场景（即，每个关系4个标签示例）中，选择在零射程设置中表现最好的LLM（text-davinci-003）。在确定在dev集上搜索的上下文示例的最优数量后，我们随机选择了具有相同实体类型约束的给定训练集中的示例。

有趣的是，普通RE无法从标记示例中获得任何改进，这表明它在少射程设置中也是有限的。普通RE表现出的限制性能表明，少射程示例可能会使模型偏向于上下文中的错误关系，而不是帮助它更准确地执行任务。

Methods	K=0	K=4	K=8	K=16	K=32
Fine-Tuning	-	9.0	21.2	29.3	33.9
PTR	-	26.8	30.0	32.9	36.8
KnowPrompt	-	30.2	33.7	34.9	35.0
$\text{NLI}_{\text{DeBERTa}}$-TEMP1	55.0	**64.2**	**64.7**	**58.7**	**65.7**
$\text{NLI}_{\text{DeBERTa}}$-TEMP2	49.4	**51.2**	47.3	**50.5**	48.1
Vanilla	48.1	46.2		-	
QA4RE	59.4	**62.0**		-	

表 4: 在 TACRED 上的少射次 F1（%）。所有结果都是对每个 K 的 3 个不同训练子集的平均值。我们对于原味 RE 和 QA4RE 使用 text-davinci-003。对于最佳表现的基线（NLI）以及原味 RE 和 QA4RE，当他们的成绩提高超过它们的零射次选择时，我们将结果标记为 **粗体**。

即使使用我们的QA4RE框架，少射击文本达芬奇-003也没有超过使用他们自己的模板(TEMP1)的DeBERTa-based NLI方法(Sainz et al., 2021)的表现。然而，即使在精心调整超参数的情况下，对NLI模型在RE数据上的微调也可能变得脆弱，这正如我们在为TEMP1和TEMP2添加更多数据时看到的不稳定性增强所证实。此外，我们发现使用TEMP2进行少射击NLI的结果从TEMP1大幅下降，这表明这种方法在少射击环境下对模板的鲁棒性也不足。因此，考虑到我们的QA方法能让LLMs通过随机选择上下文学习示例来获得少射击改进的能力，相比最佳的NLI模型，只低约2%的性能，并且对不同的模板设计具有鲁棒性，我们的方法在少射击RE上具有竞争力，且有可能通过进一步的探索实现更强的性能。我们将如何提高LLMs在少射击RE上的表现的进一步研究留给未来的工作。

Vanilla + Template RE

Given a sentence, and two entities within the sentence, classify the relationship between the two entities based on the provided sentence. All possible relationships are listed below:
- per:city_of_birth: Entity 1 was born in the city Entity 2
- per:city_of_death: Entity 1 died in the city Entity 2
- per:cities_of_residence: Entity 1 lives in the city Entity 2
- no_relation: Entity 1 has no known relations to Entity 2

Sentence: **Wearing jeans and a white blouse, Amanda Knox of Seattle is being cross-examined by prosecutors.**
Entity 1 : **Amanda Knox**
Entity 2 : **Seattle**
Relationship: **per:city_of_birth**

图 3: 与图 2中的相同例子和模板，但使用模板进行关系解释。

6 Discussions

6.1 Are Relation Templates All LLMs Need?

我们进行了消融研究，以更好地理解关系模板如何有助于QA4RE取得的性能提升。如图 3所示，我们用标记*Entity 1* 和 *Entity 2* 填充关系口译模板作为关系解释，从而将模板中的专家知识呈现给LLM。使用相同的模板和类型约束，我们将此框架（名为Vanilla+TEMP）与vanilla RE和QA4RE在TACRED数据集和GPT-3.5系列LLMs上进行比较。

如Tab. 5所示，采用同样的模板引入关系解释并不会带来一致或显著的性能提升。实际上，向任务说明中添加额外信息可能会使LLM更难理解任务。相反，使用我们的QA4RE框架，我们无需单独指定感兴趣的实体或关系解释；它们都自然地嵌入在答案选项中。这些削减结果表明，QA4RE的主要收益来自于QA的重新构造，而不仅仅来自于关系解释/模板。

6.2 QA4RE vs. NLI4RE

鉴于使用NLI重构的RE的小LMs获得的强大性能，我们利用这种相同的公式化 (Sainz et al., 2021)用于LLMs（称为NLI4RE）。[3] 更具体地说，对于每个示例，我们使用LLM预测给定的句子（前提）是否包含来自QA4RE公式的每个答案选项（假设）。我们允许LLM为每个句子-关系对生成*蕴含*，*中立*或*矛盾*。如果所有可能的正面关系中蕴含的最大概率低于0.5的阈值，则该示例将被分类为NoTA，就像Sainz et al. (2021)所做的那样。

[3]我们遵循来自ANLI的NLI格式(Wang et al., 2022).

Methods		P	R	F1	ΔF1
code-002	Vanilla	27.2	70.1	39.2	-
	Vanilla + TEMP	27.5	71.8	39.7	+0.5
	QA4RE	37.7	65.4	47.8	+8.6
text-002	Vanilla	31.2	73.1	43.7	-
	Vanilla + TEMP	26.8	77.8	39.8	−3.9
	QA4RE	35.6	68.4	46.8	+3.1
text-003	Vanilla	36.9	68.8	48.1	-
	Vanilla + TEMP	36.9	76.5	49.8	+1.7
	QA4RE	**47.7**	**78.6**	**59.4**	+11.3

表 5: 关于TACRED，评估将基于相同模板的关系解释结合到普通的RE中，是否可以弥补其与QA4RE的差距（%）。

Formulation	RED	RERED	REV	Eval	Avg.
Vanilla	48.1	55.3	51.0	36.0	47.6
NLI4RE	41.7	36.8	39.2	22.4	35.0
QA4RE	**59.4**	**61.2**	**59.4**	**43.3**	**55.8**

表 6: 文本-davinci-003的不同任务构造的F1（%）。RED, RERED, REV以及Eval分别代表TACRED, RE-TACRED, TACREV以及SemEval数据集。

如Tab. 6所示，使用NLI公式时，文本-da Vinci-003的表现出人意料地低于普通RE公式。其性能不佳的原因有两个：(1) 启发式预定义的阈值0.5对LLM来说并不理想，因此，许多正向预测被分类为NoTA。然而，在零射击设置下找到一个好的阈值也很困难。(2) 在NLI4RE下，与普通RE或QA4RE不同，LLM不是看到全部关系空间而是对每一个候选假设单独分配概率。因此，最终预测对LLM对不同关系的偏见更加敏感。

NLI4RE还需要对每个关系实例进行多次推理运行以评估所有候选关系，这导致了显着更高的成本。

6.3 QA4RE & Model Size

为了验证我们的QA4RE框架在较小的指令调优模型上的有效性和可转移性，我们进一步在四个RE数据集的完整测试集上评估了FLAN-

LMs	Model Size	Avg. F1		
		Vanilla	QA4RE	Δ
GPT-3.5 Series				
text-001	175B	22.3	14.9	−7.4
code-002	175B	39.8	43.9	+4.1
text-002	175B	43.2	45.9	+2.7
text-003	175B	47.6	55.8	+8.2
FLAN-T5 Series				
Small	80M	19.5	25.0	+5.6
Base	250M	22.3	26.4	+4.2
Large	780M	34.8	41.8	+7.0
XLarge	3B	46.6	54.3	+7.7
XXLarge	11B	46.5	55.1	+8.6

表 7: QA4RE在GPT-3.5系列和不同大小的FLAN-T5上的有效性。结果是对四个RE数据集的平均值。

T5 Small (80M)，Base (250M)，和 Large (780M)。表 7 展示了我们的QA4RE框架仍然可以为各种大小的指令调优模型带来相当大的提升，即使是对于最小的一个（80M）。这证明了QA4RE的有效性可以在从80M到175B的各种模型大小中进行转移，考虑到QA4RE在几个GPT-3.5模型上持续的提升。

在FLAN-T5系列中，较大的模型更能从我们的框架中受益。然而，我们注意到，当扩大到更大的GPT-3.5模型时，这一趋势并未继续。事实上，除了text-davinci-003外，所有的GPT-3.5模型从QA4RE中受益的程度都比FLAN-T5模型少。 QA4RE在这些模型上的改进较小，使得它们的总体RE性能只与大约小20倍和50倍的模型相当。这表明即使在与高发病率任务对齐时，Sec. 2 中所讨论的GPT-3.5系列模型采用的各种对齐策略可能并不比标准指令调优对于提高模型在低发病率任务上的泛化性更有效。然而，在测试中最强的模型，text-davinci-003和FLAN-T5-XXL，观察到的强烈改进显示出QA4RE的有效性在未来模型变得更加能力强大时有可能继续保持。

7 Conclusions and Future Work

在这项工作中，我们首先展示了即使是最新的经过指令调优的LLMs在关系抽取（RE）任务上也不如细调的小型LMs的表现。为了解辟这个限制，我们将RE重新构想为多项选择题答题（QA），目的是利用一项在像QA这样的指令调优数据集中广泛覆盖的任务，而不是RE，RE在这些数据集中几乎不存在。全面的实验表明，我们的QA4RE框架让LLMs展现出了作为零射击（zero-shot）关系抽取器的力量，特别是对于两个最新的LLMs（text-davinci-003和FLAN-T5 XXL）。我们还进行了深入的实验，以探索我们的方法的稳健性和少射击（few-shot）有效性，以及研究在哪些LLM训练场景中，它是最有效的。

在未来的工作中，我们希望探索在指令调优中可能对LLMs构成挑战、且可能成功对齐更广泛采用的指令调优任务（例如QA）的其他代表性不足的任务。此外，我们计划继续探索这条研究路径，通过运用我们的QA4RE框架为其他LLMs提供支持，比如OPT系列 (Zhang et al., 2022; Iyer et al., 2022)和PaLM (Chowdhery et al., 2022)，由于受限于计算资源和/或访问权限，这些在本工作中并未包括进来。

8 Limitations

尽管我们的方法有助于释放六种最新强大的LLMs作为零射程关系提取器的能力，但早期的LLM例如没有进行强大的指令调整的text-davinci-001并未从我们的框架中看到任何改进。此外，虽然我们在零射程RE设置上进行了全面的实验，但我们的少射程探索则更为有限。从我们的研究中还不清楚，是否包含更多训练样例能否提高LLM的RE性能，以及在少射程设置中，是否会延续在GPT-3模型零射程设置中观察到的相同趋势。我们将这些问题的回答留给未来的工作。

9 Ethics Statement

在这项工作中，我们提出了一种方法以提高LLM在关系提取这项重要且基础任务上的表现。我们并不预见到关于这项研究主题的任何道德问题。

Acknowledgements

作者感谢娄仁泽，OSU NLP小组的同事们以及匿名审稿人对他们宝贵的反馈表示感谢。作者也要感谢陆克明关于再现SuRE的讨论和指导。本研究部分由NSF OAC 2112606，NIH R01LM014199，和俄亥俄州超级计算中心支持(Center, 1987)。

References

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ohio Supercomputer Center. 1987. Ohio supercomputer center.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2778–2788. ACM.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Geli Fei and Bing Liu. 2016. Breaking the closed world assumption in text classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 506–514. The Association for Computational Linguistics.

Hao Fu, Yao; Peng and Tushar Khot. 2022. How does gpt obtain its ability? tracing emergent abilities of language models to their sources. *Yao Fu' s Notion*.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. Fewrel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 6249–6254. Association for Computational Linguistics.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2022. Ptr: Prompt tuning with rules for text classification. *AI Open*, 3:182–192.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2022. OPT-IML: scaling language model instruction meta learning through the lens of generalization. *CoRR*, abs/2212.12017.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 in-context learning for biomedical IE? think again. In *Findings of EMNLP*, pages 4497–4512, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher R'e, Diana Acosta-Navas, Drew A. Hudson, E. Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan S. Kim, Neel Guha, Niladri S. Chatterji, O. Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas F. Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models. *ArXiv*, abs/2211.09110.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022, Dublin, Ireland and Online, May 27, 2022*, pages 100–114. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, and Muhao Chen. 2022. Summarization as indirect supervision for relation extraction. In *Findings of EMNLP*, pages 6575–6594, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022a. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5316–5330. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022b. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True Few-Shot Learning with Language Models.

Ofer Sabo, Yanai Elazar, Yoav Goldberg, and Ido Dagan. 2021. Revisiting few-shot relation classification:

Evaluation data and classification schemes. *Trans. Assoc. Comput. Linguistics*, 9:691–706.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of EMNLP*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Lei Shu, Hu Xu, and Bing Liu. 2017. DOC: deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2911–2916. Association for Computational Linguistics.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the TACRED dataset. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13843–13850. AAAI Press.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions:generalization via declarative instructions on 1600+ tasks. In *EMNLP*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 35–45. Association for Computational Linguistics.

A Instruction Dataset Portion

	#Tasks	%RE	%QA
T0 (Sanh et al., 2022)	62	0	27.4
FLAN (Wei et al., 2022a)	62	0	21
MetaICL (Min et al., 2022b)	142	0	28.9
NaturalInstruct (Wang et al., 2022)	1731	<0.5	>12

表 9: 流行的指令调优数据集以及每个数据集中RE和QA任务的比例。

如Tab. 9所示，T0 (Sanh et al., 2022)，FLAN (Wei et al., 2022a)和MetaICL (Min et al., 2022b)的指令调优数据集中没有RE任务。即使在目前最大的NaturalInstruct (Wang et al., 2022)中，RE任务也只占总任务量的不到0.5%。相比之下，QA是所有指令调优数据集中最常见的任务格式。这些观察结果表明，用于指令调整的数据集中RE任务的发生频率较低，而QA任务则占据主导地位。

B Experimental Details

B.1 Hyperparameters for Few-Shot Methods

在少样本设置下，对于每个K，我们随机抽样 3 次得到不同的训练子集，每个子集会作为LLM的上下文示例或用于训练基线中的小型语言模型。报告结果是对三个子集求平均。为避免由于使用过多的 dev 示例而高估了少样本性能 (Perez et al., 2021)，我们为所有的超参数搜索使用 100 个随机选择的 dev 集示例。

对于LLMs，我们使用 dev 集来搜索来自 $\{1, 2, 5\}$ 的最优数量的上下文示例作为一个超参数。然后我们从给定的训练集中随机选择相同类型的受限制的上下文示例。

对于所有基于小型 LM 的基线，我们使用他们公开提供的代码和用于训练的超参数。根据NLI (Sainz et al., 2021)和SuRE (Lu et al., 2022)的原始论文，我们使用在线提供的检查点和报告的超参数进行模型训练。遗憾的是，我们无法使用默认超参数复现SuRE的结果。对于标准的Fine-Tuning (Jimenez Gutierrez et al., 2022)，PTR (Han et al., 2022)，和KnowPrompt (Chen et al., 2022)，我们在开发环境上使用Tab. 10所示的范围对超参数进行网格搜索。

我们使用8片NVIDIA GeForce RTX 2080 Ti和2片NVIDIA RTX A6000来进行所有实验。总共使用的GPU小时数和OpenAI API的费用列在Tab. 11中。

Hyperparameter	Search Space
Learning Rate 1:	{1e−5, 3e−5}
Weight Decay:	{0.01, 0.001}
Learning Rate 2:	{5e−5, 2e−4}

表 10: 用于少镜像方法的网格搜索的超参数。学习率2用于PTR (Han et al., 2022)的新令牌训练和KnowPrompt (Chen et al., 2022)的虚拟令牌训练。

	Num of Params (Millions)	Total GPU Hours	Total Cost
RoBERTa-Large	354	284	-
DeBERTa-XLarge	900	14	-
BART-Large	406	2	-
Pegasus-Large	568	50	-
FLAN-T5 S	80	<1	-
FLAN-T5 M	250	<1	-
FLAN-T5 L	780	1	-
FLAN-T5 XL	3,000	2	-
FLAN-T5 XXL	11,000	4	-
OpenAI Text API	175,000	-	$835
OpenAI Chat API	?	-	$4

表 11: 对于开源LMs的总GPU小时数和使用OpenAI API的费用（包括所有版本）。

B.2 Prompts for LLMs

如Tab. 12所示，我们列出了本文中使用的所有模板，包括在Tab. 5中的vanilla + TEMP，在Tab. 6中的NLI4RE，以及在所有实验中的vanilla以及QA4RE。

B.3 Relation Verbalization Templates

在 Tab. 2 显示的关系 verbalization 模板的鲁棒性实验中，以下使用来自 TACRED 基准的 *org:top_members/employees* 关系作为例子描述了四个模板之间的差异：

1. 具体示例: $\{E_h\}$是$\{E_t\}$的主席/总统/董事长
2. 语义关系：$\{E_h\}$是$\{E_t\}$的高级成员

Methods		TACRED			RETACRED			TACREV			SemEval			Avg.
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	F1
Small	Vanilla	9.5	40.9	15.4	22.8	50.2	31.3	9.1	41.9	15.0	10.0	11.8	10.8	18.1
	QA4RE	13.8	52.2	21.8 (+6.4)	33.5	66.2	44.5 (+13.2)	13.7	55.2	22.0 (+7.0)	5.9	7.1	6.4 (−4.4)	23.7 (+5.6)
Base	Vanilla	14.1	31.1	19.4	21.1	26.8	23.6	14.1	33.3	19.8	14.9	17.9	16.2	19.8
	QA4RE	17.1	54.7	26.0 (+6.6)	33.0	65.2	43.8 (+20.2)	17.2	58.5	26.6 (+6.8)	6.7	8.0	7.3 (−8.9)	25.9 (+6.2)
Large	Vanilla	22.8	58.6	32.8	37.5	60.8	46.4	22.6	61.9	33.1	23.7	19.7	21.5	33.5
	QA4RE	30.3	78.5	43.7 (+10.9)	44.5	72.6	55.2 (+8.8)	29.9	82.4	43.9 (+10.8)	24.8	15.8	19.3 (−2.2)	40.5 (+7.1)
XLarge	Vanilla	48.8	49.0	48.9	55.8	39.8	46.4	52.0	55.7	**53.8**	34.9	29.6	32.0	45.3
	QA4RE	37.6	78.6	50.9 (+2.0)	56.2	79.9	66.0 (+19.6)	38.2	84.7	52.7 (−1.1)	44.4	39.9	42.1 (+10.1)	52.9 (+7.7)
XXLarge	Vanilla	48.2	45.3	46.7	56.1	53.7	54.9	50.6	50.6	50.6	29.2	28.1	28.6	45.2
	QA4RE	38.1	82.9	**52.2** (+5.5)	55.9	82.0	**66.5** (+11.6)	38.3	88.1	53.4 (+2.8)	40.2	47.5	**43.5** (+14.9)	**53.9** (+8.7)

表 8: FLAN-T5在四个RE数据集的完整测试集上的结果（%）。我们将最好的结果标注为**粗体**，第二好的结果下划线，并将我们的QA4RE相比于普通RE的F1改进标记为绿色。

3. 直白：集合$\{E_h\}$和集合$\{E_t\}$之间的关系是顶级成员或员工

4. 词语翻译：$\{E_h\}$ 组织级别的成员或雇员 $\{E_t\}$

第一套模板由Sainz et al. (2021)编写，而其余三套则由Lu et al. (2022)探索。我们从他们的官方GitHub仓库中使用模板。[4] 此外，我们还在Tab. 13、Tab. 14 和 Tab. 15中列出了我们的论文中所有LLMs使用的关系语义化模板。

C Full Test Results on FLAN-T5

我们在表 8中呈现了所有四个RE数据集的完整测试集结果。我们的观察结果与在1,000个测试示例上的实验结果相符：

(1) 我们的QA4RE框架可以在所有FLAN-T5系列模型的平均结果上带来一致且显著的提升。此外，更大的模型更能从我们的框架中获益。这两个信号强烈证明了QA4RE的有效性。

(2) 我们注意到，我们的QA4RE并未改善在SemEval上的FLAN-T5的小版本，这是一个19选项的QA任务。这可能是由于这些模型在理解由QA4RE提供的带长输入时遇到困难。

[4]用于健壮性实验的模板:
TEMP1：https://github.com/osainz59/Ask2Transformers/blob/master/resources/predefined_configs/tacred.relation.config.json
TEMP3：https://github.com/luka-group/SuRE/blob/main/data templates/tacred/rel2temp_forward.json
TEMP4：https://github.com/luka-group/SuRE/blob/main/data/templates/tacred/rel2temp_raw_relation.json

Formulations	Prompts
Vanilla RE	Given a sentence, and two entities within the sentence, classify the relationship between the two entities based on the provided sentence. All possible Relationships are listed below: - [Possible Relation 1] - [Possible Relation 2] - [NoTA Relation] Sentence: [Sentence S] Entity 1: [Head Entity E_h] Entity 2: [Tail Entity E_t] Relationship:
Vanilla + TEMP	Given a sentence, and two entities within the sentence, classify the relationship between the two entities based on the provided sentence. All possible Relationships are listed below with explanations: - [Possible Relation 1]: [Relation 1 Template] - [Possible Relation 2]: [Relation 2 Template] - [NoTA Relation]: [NoTA Relation Template] Sentence: [Sentence S] Entity 1: [Head Entity E_h] Entity 2: [Tail Entity E_t] Relationship:
NLI4RE	In this task, you will be presented with a premise and a hypothesis sentence. Determine whether the hypothesis sentence entails (implies), contradicts (opposes), or is neutral with respect to the given premise sentence. Please answer with "Contradiction", "Neutral", or "Entailment". Premise: [Sentence S] Hypothesis: [Entities in Relation 1 Template] Category:
QA4RE	Determine which option can be inferred from the given Sentence. Sentence: [Sentence S] Options: A. [Entities in Relation 1 Template] B. [Entities in Relation 2 Template] C. [Entities in NoTA Relation Template] Which option can be inferred from the given Sentence? Option:

表 12: 本文中的LLMs框架的格式提示。为了简单起见，我们只用1个模板展示NLI4RE。

Relation	**Template**
no_relation	$\{E_h\}$ has no known relations to $\{E_t\}$
per:stateorprovince_of_death	$\{E_h\}$ died in the state or province $\{E_t\}$
per:title	$\{E_h\}$ is a $\{E_t\}$
org:member_of	$\{E_h\}$ is the member of $\{E_t\}$
per:other_family	$\{E_h\}$ is the other family member of $\{E_t\}$
org:country_of_headquarters	$\{E_h\}$ has a headquarter in the country $\{E_t\}$
org:parents	$\{E_h\}$ has the parent company $\{E_t\}$
per:stateorprovince_of_birth	$\{E_h\}$ was born in the state or province $\{E_t\}$
per:spouse	$\{E_h\}$ is the spouse of $\{E_t\}$
per:origin	$\{E_h\}$ has the nationality $\{E_t\}$
per:date_of_birth	$\{E_h\}$ has birthday on $\{E_t\}$
per:schools_attended	$\{E_h\}$ studied in $\{E_t\}$
org:members	$\{E_h\}$ has the member $\{E_t\}$
org:founded	$\{E_h\}$ was founded in $\{E_t\}$
per:stateorprovinces_of_residence	$\{E_h\}$ lives in the state or province $\{E_t\}$
per:date_of_death	$\{E_h\}$ died in the date $\{E_t\}$
org:shareholders	$\{E_h\}$ has shares hold in $\{E_t\}$
org:website	$\{E_h\}$ has the website $\{E_t\}$
org:subsidiaries	$\{E_h\}$ owns $\{E_t\}$
per:charges	$\{E_h\}$ is convicted of $\{E_t\}$
org:dissolved	$\{E_h\}$ dissolved in $\{E_t\}$
org:stateorprovince_of_headquarters	$\{E_h\}$ has a headquarter in the state or province $\{E_t\}$
per:country_of_birth	$\{E_h\}$ was born in the country $\{E_t\}$
per:siblings	$\{E_h\}$ is the siblings of $\{E_t\}$
org:top_members/employees	$\{E_h\}$ has the high level member $\{E_t\}$
per:cause_of_death	$\{E_h\}$ died because of $\{E_t\}$
per:alternate_names	$\{E_h\}$ has the alternate name $\{E_t\}$
org:number_of_employees/members	$\{E_h\}$ has the number of employees $\{E_t\}$
per:cities_of_residence	$\{E_h\}$ lives in the city $\{E_t\}$
org:city_of_headquarters	$\{E_h\}$ has a headquarter in the city $\{E_t\}$
per:children	$\{E_h\}$ is the parent of $\{E_t\}$
per:employee_of	$\{E_h\}$ is the employee of $\{E_t\}$
org:political/religious_affiliation	$\{E_h\}$ has political affiliation with $\{E_t\}$
per:parents	$\{E_h\}$ has the parent $\{E_t\}$
per:city_of_birth	$\{E_h\}$ was born in the city $\{E_t\}$
per:age	$\{E_h\}$ has the age $\{E_t\}$
per:countries_of_residence	$\{E_h\}$ lives in the country $\{E_t\}$
org:alternate_names	$\{E_h\}$ is also known as $\{E_t\}$
per:religion	$\{E_h\}$ has the religion $\{E_t\}$
per:city_of_death	$\{E_h\}$ died in the city $\{E_t\}$
per:country_of_death	$\{E_h\}$ died in the country $\{E_t\}$
org:founded_by	$\{E_h\}$ was founded by $\{E_t\}$

表 13: Templates for TACRED and TACREV datasets.

Relation	Template
no_relation	$\{E_h\}$ has no known relations to $\{E_t\}$
per:religion	$\{E_h\}$ has the religion $\{E_t\}$
org:country_of_branch	$\{E_h\}$ has a branch in the country $\{E_t\}$
org:stateorprovince_of_branch	$\{E_h\}$ has a branch in the state or province $\{E_t\}$
org:city_of_branch	$\{E_h\}$ has a branch in the city $\{E_t\}$
org:shareholders	$\{E_h\}$ has shares hold in $\{E_t\}$
org:top_members/employees	$\{E_h\}$ has the high level member $\{E_t\}$
org:members	$\{E_h\}$ has the member $\{E_t\}$
org:website	$\{E_h\}$ has the website $\{E_t\}$
per:parents	$\{E_h\}$ has the parent $\{E_t\}$
org:number_of_employees/members	$\{E_h\}$ has the number of employees $\{E_t\}$
org:political/religious_affiliation	$\{E_h\}$ has political affiliation with $\{E_t\}$
per:age	$\{E_h\}$ has the age $\{E_t\}$
per:origin	$\{E_h\}$ has the nationality $\{E_t\}$
org:alternate_names	$\{E_h\}$ is also known as $\{E_t\}$
per:other_family	$\{E_h\}$ is the other family member of $\{E_t\}$
per:identity	$\{E_h\}$ is the identity/pronoun of $\{E_t\}$
per:identity	$\{E_h\}$ and $\{E_t\}$ are the same person
per:siblings	$\{E_h\}$ is the siblings of $\{E_t\}$
org:member_of	$\{E_h\}$ is the member of $\{E_t\}$
per:children	$\{E_h\}$ is the parent of $\{E_t\}$
per:employee_of	$\{E_h\}$ is the employee of $\{E_t\}$
per:spouse	$\{E_h\}$ is the spouse of $\{E_t\}$
org:dissolved	$\{E_h\}$ dissolved in $\{E_t\}$
per:schools_attended	$\{E_h\}$ studied in $\{E_t\}$
per:country_of_death	$\{E_h\}$ died in the country $\{E_t\}$
per:stateorprovince_of_death	$\{E_h\}$ died in the state or province $\{E_t\}$
per:city_of_death	$\{E_h\}$ died in the city $\{E_t\}$
per:date_of_death	$\{E_h\}$ died in the date $\{E_t\}$
per:cause_of_death	$\{E_h\}$ died because of $\{E_t\}$
org:founded	$\{E_h\}$ was founded in $\{E_t\}$
org:founded_by	$\{E_h\}$ was founded by $\{E_t\}$
per:countries_of_residence	$\{E_h\}$ lives in the country $\{E_t\}$
per:stateorprovinces_of_residence	$\{E_h\}$ lives in the state or province $\{E_t\}$
per:cities_of_residence	$\{E_h\}$ lives in the city $\{E_t\}$
per:country_of_birth	$\{E_h\}$ was born in the country $\{E_t\}$
per:stateorprovince_of_birth	$\{E_h\}$ was born in the state or province $\{E_t\}$
per:city_of_birth	$\{E_h\}$ was born in the city $\{E_t\}$
per:date_of_birth	$\{E_h\}$ has birthday on $\{E_t\}$
per:charges	$\{E_h\}$ is convicted of $\{E_t\}$
per:title	$\{E_h\}$ is a $\{E_t\}$

表 14: Templates for RETACRED datasets.

Relation	Template
Other	{subj} has no known relations to {obj}
Component-Whole(e1,e2)	{subj} is the component of {obj}
Component-Whole(e2,e1)	{obj} is the component of {subj}
Instrument-Agency(e1,e2)	{subj} is the instrument of {obj}
Instrument-Agency(e2,e1)	{obj} is the instrument of {subj}
Member-Collection(e1,e2)	{subj} is the member of {obj}
Member-Collection(e2,e1)	{obj} is the member of {subj}
Cause-Effect(e1,e2)	{subj} has the effect {obj}
Cause-Effect(e2,e1)	{obj} has the effect {subj}
Entity-Destination(e1,e2)	{obj} is the destination of {subj}
Entity-Destination(e2,e1)	{subj} is the destination of {obj}
Content-Container(e1,e2)	{obj} contains {subj}
Content-Container(e2,e1)	{subj} contains {obj}
Message-Topic(e1,e2)	{obj} is the topic of {subj}
Message-Topic(e2,e1)	{subj} is the topic of {obj}
Product-Producer(e1,e2)	{obj} produces {subj}
Product-Producer(e2,e1)	{subj} produces {obj}
Entity-Origin(e1,e2)	{subj} origins from {obj}
Entity-Origin(e2,e1)	{obj} origins from {subj}

表 15: Templates for SemEval datasets.