声明：我已知悉学校对于考试纪律的严肃规定，将秉持诚实守信宗旨，严守考试纪律，不作弊，不剽窃；若有违反学校考试纪律的行为，自愿接受学校严肃处理。

2024-2025 学年第一学期 COMP737026 《大模型理论与方法》课程期末论文

学号：24210240291 , 姓名：宋安洋, 贡献：50%
学号：24210240386 , 姓名：于洪洲, 贡献：50%

Abstract

我们的期末作业题目是：探索各种 ICL 方法在 NLP Beginner 中的应用。task1/task2 是 kaggle 影评片段的情感五分类。task3 是斯坦福语义推理的三分类。task4 是命名实体识别，九个类别的序列标注问题。task5 是中文唐诗生成。我们在完成 NLP Beginner 使用 CNN、LSTM 的原任务，并实现了 BERT 类模型用作对比的基础上，考虑到 task4 可能并不是很适合使用纯的 ICL，以及 task5 没有对比指标，我们对 task1/2、task3 尝试了一些 ICL 的方法并进行对比。其中，组员的贡献如下：（1）宋安洋：完成了指令学习和上下文学习的综述。完成了 NLP Beginner 原任务与改用 BERT 类模型的实验。完成了对 ICL 模板的调整与尝试。完成了 Channel、Calibrate、KATE 的复现与对比，并完成了对应部分实验报告的撰写。（2）于洪洲：完成了自然语言推理和命名实体识别的综述。完成了 ICL 不同 shot 数的对比实验，并完成了对应部分实验报告的撰写。我们的实验贡献如下：（1）证明了 ICL 对模板格式和长度的高度敏感，探索了尽可能最佳的模板格式。（2）证明了未经过指令微调的大模型在面对困难的特定分类任务时表现远不如训练的分类器。ICL 更适合与其他方法结合来取得更好的表现。（3）证明了相比分类器，ICL 的最大优点是不存在过拟合问题，拥有强大的泛化能力。（4）发现了生成模型中的自然语言偏见某些情形下可能是必要的。（5）证明了 Channel 方法本身就是几乎不受偏见影响的，与偏见几乎无关。我们的代码在：https://github.com/s1012480564/NLP_Beginner_new 可以获取到。

1 引言

人工智能的目标之一是建立一个能够普遍理解和解决新任务的系统。目前，为样本打标签是主流的任务表示方式，然而大规模获取标签的成本很高，甚至在某些情况下难以进行注释或打标签，或者甚至根本做不到的，没有恰当的打标签的方式。在这种情况下，文本指令为表达任务语义提供了另一个监督维度，它通常包含比单个标记示例更抽象和更全面的目标任务知识。有了任务指令，就可以快速构建系统来处理新任务。在现实应用中，这种效率是非常理想的，特别是当特定于任务的注释稀缺时。更重要的是，在学习新任务方面，遵循指令倾向于人类智能，就好比一个小孩子可以通过学习指令和一些例子来很好地解决新的数学任务 [32, 11]。因此，这种新的学习范式近年引起了 NLP 领域的重要关注 [133, 77]。说起指令学习，首先会想到“提示”——比如使用一个简短的模板将任务输入转换为新的格式（例如完形填空问题），以满足大型语言模型（LLM）的语言建模目标 [8]。尽管提示在文本分类、机器翻译等领域很普遍，但它只是指令的一种特殊情况。而近两年指令学习中最热门的一种特殊形式是，上下文学习，或称作情境学习（In-Context Learning, ICL）。

随着模型大小和数据大小的扩展 [9, 18, 94, 126, 127]，大型语言模型（LLM）展示了上下文学习（ICL）的能力，即从上下文中的几个示例（demonstration Examples）中学习。许多研究表明，LLM 可以通过 ICL 执行一系列复杂的任务，例如解决数学推理问题 [138]。这些强

大的能力已被广泛验证为大型语言模型的新兴能力 [137]。ICL 的关键思想是从类比中学习：首先，ICL 需要一些示例来形成提示上下文。这些例子通常用自然语言模板编写。然后，ICL 将查询问题和提示上下文连接在一起形成输入，然后将其输入到语言模型中进行预测。与需要训练、使用反向传播的梯度来更新模型参数的监督学习不同，ICL 不执行参数更新。ICL 能够让模型学习示例中隐藏的模式，并相应地做出正确的预测。作为一种新的范式，ICL 具有多种吸引人的优势。第一，由于示例是用自然语言编写的，因此它提供了一个可解释的界面来与 LLM 进行交流 [9]。这种范式使得通过改变示例和模板将人类知识融入 LLM 变得更加容易 [70, 79, 138, 145]。第二，ICL 类似于人类通过类比学习的决策过程 [143]。第三，与监督训练相比，ICL 是一种免训练的学习框架。这不仅可以大大降低模型适应新任务的计算成本，还可以使语言模型即时服务 [120] 成为可能，并且可以轻松应用于大规模的现实世界任务。尽管前景广阔，但 ICL 还存在一些有趣的问题和有趣的特性需要进一步研究。本文中，通过实验，尝试了将 ICL 应用于 NLP Beginner，并基于实验结果分析探讨了 ICL 的一些有趣的特性和问题。关于自然语言推理等任务，对人类来说很自然，但往往对机器来说很困难，如果能够让大语言模型用极其自然的自然语言生成的方式生成出准确的答案，那将是极其令人兴奋的。

对于自然语言推理（NLI），用一个更直观的说法来解释是，我们人类使用各种知识和对文本本身的理解来推理语言背后隐藏的含义。例如，考虑 [87] 中的句子：“Jack needed some money, so he went and shook his piggy bank. He was disappointed when it made no sound.”由此我们就不难理解，杰克没有找到钱，也正因为如此，杰克才产生了负面情绪。使我们得出这个结论的原因是我们对世界的了解以及潜在的推理过程，通常被称为常识思维 [87] 或常识推理 [24] ，这使我们在文章中没有明确说明这一点的前提下，能够将知识片段连接起来以得出新的结论。但是，虽然这种知识和推理对于人类读者来说是很自然的，但对于机器来说却是出了名的困难。尽管过去几十年来自然语言处理取得了重大进展，但机器距离具备这种自然语言推理（NLI）能力还很远。对于相关进展，本文重点总结了斯坦福语义推理（SNLI）任务中的一些模型的优化与改进。

另一个至关重要的任务是命名实体识别（NER）。目前，数据通常以原始形式呈现（非结构化、母语、混乱），来自我们经济、政府、私人和公共生活的各个部门，因此对于人类来说，总结、搜索、得出结论和进行统计分析都是具有挑战性的任务。这使得 NLP 在当今的数据处理中至关重要。其中一项操作便是命名实体识别，对文本文档中呈现的命名实体进行识别和分类的任务称为命名实体识别。1997 年， Chinchor and Robinson [16] 在第六届消息理解会议 (MUC) 会议上首次使用“命名实体识别”一词来识别所有表达式。NER 在各个不同领域都有广泛的应用。比如，社交媒体 [58] 每天产生大量数据。而社交媒体写作并不严格遵守句法原则，这导致语言背后蕴含的关键信息往往隐藏在组织松散的社交媒体帖子的形式中。因此，NER 是一个至关重要的确定文本中的适当实体并为后续 NLP 活动如情感分析提供帮助的步骤。

本文中，我们首先对指令学习进行了进一步的、更全面、广泛的总结和探讨。我们尝试回答以下问题：（1）什么是任务指令，存在哪些指令类型？（2）给定任务指令，如何对其进行编码以协助目标任务的模型泛化？（3）哪些因素（例如模型大小、任务数量）影响指令驱动系统的性能？（4）指令学习存在哪些挑战，未来的方向是什么？然后我们对上下文学习的主要研究方面进行了总结和分析，包括：（1）我们总结了通过预训练期间的适应显着提高 ICL 能力的方法。（2）ICL 的性能对特定设置很敏感，我们总结了包括提示模板、演示示例的选择和顺序以及其他因素。（3）我们总结并指出了，尽管有初步的解释，ICL 的潜在工作机制仍不清楚，需要进一步研究。接着，我们对斯坦福自然语言推理和命名实体识别中的一些方法进行了总结。最后我们完成了将 ICL 应用于 NLP Beginner 并与 CNN、LSTM、BERT 等训练方法的对比实验，对实验结果进行了详细的分析并得到了一些重要的结论。

2 概念定义

本节中，我们将分别对指令学习、上下文学习、自然语言推理、命名实体识别这四个领域及领域内的相关概念给出更加明确的定义。

2.1 指令学习

对于指令学习（Instruction Learning），我们的目标是通过学习指令来驱动系统达到输入的相应输出。因此，我们假设数据集通常包括三项：

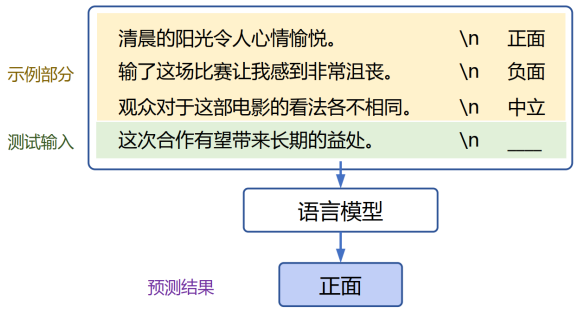

Figure 1: 上下文学习（ICL）的图示。ICL 需要一个上下文的提示词，其中包含一些用自然语言模板编写的示例（demonstration）。将这种提示词和查询作为输入，大型语言模型负责进行预测。

- **输入** (X)：实例的输入；它可以是单个文本片段（例如，情感分类任务）或一组文本片段（例如，文本推理、文本 QA 等）。
- **输出** (Y)：实例的输出；在分类问题中，它可以是一个或多个预定义的标签；在文本生成任务中，它可以是任何开放形式的文本。
- **模板** (T)：试图表达任务意图或用于连接 X 和 Y。T 的文本模板本身往往可能还并不是指令。

我们将在第3.1.1节中详细说明，任务指令 I 是实际上是 T 与 X 或 Y 的组合，或者在某些情况下是 T 本身。

2.2 上下文学习

上下文学习（In-Context Learning, ICL）可以认为是指令学习的一个子类，也可以将其与常规的指令区分来看待。基于 Brown et al. [9]，我们在可以给出上下文学习的一个正式定义：

> 上下文学习是一种在仅以演示的形式给出几个例子的前提下允许语言模型学习任务的范式。

形式上，在图 1 中给出了一个简单直观的例子。给定一个查询输入文本 x 和一组候选答案 $Y = \{y_1, \ldots, y_m\}$，预先训练的语言模型 $\mathcal{M}$ 基于条件演示集 C，将得分最高的候选答案作为预测。C 包含一个可选任务指令 I 和 k 个示例（demonstration），因此 $C = \{I, s(x_1, y_1), \ldots, s(x_k, y_k)\}$ 或 $C = \{s'(x_1, y_1, I), \ldots, s'(x_k, y_k, I)\}$，其中 $s'(x_i, y_i, I)$ 是根据任务用自然语言写的一个例子。根据 k 和示例是否属于同一任务，可以将其分为任务特定的 ICL 和跨任务的 ICL。在后一种情况下，不同的例子有自己的说明。候选答案 y_j 的可能性来自整个输入序列上的计分函数

$$P(y_j \mid x) \triangleq f_{\mathcal{M}}(y_j, C, x) \tag{1}$$

最终预测标签 $\hat{y}$ 是具有最高概率的候选答案：

$$\hat{y} = \arg\max_{y_j \in Y} P(y_j \mid x) \tag{2}$$

根据 ICL 的定义，我们可以看到 ICL 与相关概念的区别如下：(1) 提示学习（Prompt Learning）：ICL 属于提示学习的一个子类，但和一般的提示学习的区别在于 ICL 引入了示例作为提示词的一部分。例如 Liu et al. [72]对提示学习做了详细彻底的综述，但 ICL 不在他们的研究范围内。(2) 小样本学习（Few-shot Learning）：小样本学习是一种通用的机器学习方法，它涉及调整模型参数来执行监督样本数量有限的任务 [132]。ICL 属于小样本学习的一个子类，但和一般的小样本学习相反的是，ICL 不需要参数更新，而是直接在预先训练的 LLM 上执行。

2.3 自然语言推理

自然语言推理（Natural Language Inference, NLI），也称为识别文本蕴涵（RTE, Recognizing Textual Entailment），是确定两个（简短的、有序的）文本之间的推理关系的任务：蕴涵

前提	推理关系	假说
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

Table 1: 斯坦福语义推理（SNLI）的示例样本。(其中每行 5 个的 E、C、N 是标注者的投票，由多数投票结果决定推理关系标签。可以忽略)

(entailment)、矛盾（contradiction）或中性（neutral） [81]。蕴涵即假说包含在前提中，或者说前提能够推理出假设。矛盾即假说和前提相反，或者说前提能够推理出假说不成立。中性即假说和前提不相关，即无法构成推理关系，不知道假说成立还是不成立。表1中展示了斯坦福语义推理（SNLI）的一些示例样本。

2.4 命名实体识别

命名实体识别（Named Entity Recognition, NER）是识别文本中引用的大量信息片段，然后将它们分类到预先建立的类别的过程。“人”、“组织”、“区域”等实体可以被视为类别。命名实体识别作为序列标注任务是 NLP 问题中的一个大类。NER 之外的 NLP 序列标注任务包括分块和词性标注。当模型用于序列标注时，会为单词或标记序列中出现的每个单词生成标签。问答、信息检索、机器翻译和所有其他应用程序都受益于命名实体识别提供的预处理。一个 NER 系统是，给定一个序列的元组 ($s = \langle w_1 \rangle, \langle w_2 \rangle, ..., \langle w_N \rangle$); 它将输出一组标签 ($\langle L_1 \rangle, \langle L_2 \rangle, ..., \langle L_N \rangle$) 。

3 文献综述

3.1 指令学习

3.1.1 任务的指令是什么？

从指令格式的角度来看包括三种类型:

(1) 面向 NLI 的指令（即 $I = T + Y$): 处理分类任务的传统方案是将目标标签转换为索引，然后让模型决定输入属于哪个索引。这种范式仅对输入语义进行编码，而丢失了标签语义。为了让系统在不依赖大量标记示例的情况下识别新标签，Yin et al. [159] 提出通过为每个标签建立假设，将目标分类任务转换为自然语言推理（NLI）——导出标签的真值，然后转换为确定假设的真值。它已广泛应用于各种少样本/零样本分类任务 [150]，例如主题分类 [159]、情感分类 [167]、立场分类 [151]、实体类型 [61]、实体关系 [90, 146, 108, 109]。

(2) 面向 LLM 的指令（即提示词、$I = T + X$): 提示是面向 LLM 的指令的代表，通常是前面加上任务输入的简短话语（前缀提示），或完形填空问题模板（完形填空提示）。它基本上是为了查询来自 LLM 中间媒体的响应（可以进一步转换为最终输出）而设计的。由于提示输入符合 LLM 的预训练目标（例如，完形填空输入满足屏蔽语言建模目标 [27]，因此有助于摆脱对传统监督微调的依赖，并大大减轻人工注释的成本。因此，提示学习在许多先前的少数/零样本 NLP 任务上取得了令人印象深刻的结果，例如问答 [105, 69]、机器翻译 [66]、情感分析 [144]、文本蕴涵 [110, 111]、实体识别 [22, 129]。

(3) 人性化指令（即 $I = T + optional\{X_i, Y_i\}_{i=1}^{k}$): 与面向 LLM 的指令不同，面向人的指令（图 2（c)）通常是一些人类可读的、描述性的、段落式的信息，由各种组件组成，例如

“任务标题”、“类别”、“定义”和“要避免的事情”等 [89]。因此，面向人类的指令更加用户友好，可以理想地应用于几乎任何复杂的 NLP 任务。因此，面向人类的指令近年来引起了更多关注 [46, 40, 160]。但由于其复杂性，面向人类的指令更具挑战性。

从间接监督的角度来看：指令学习是，虽然上述三种类型的指令彼此有很大不同，但它们本质上都在寻求相同的东西——间接监督——以应对注释有限的目标任务。

3.1.2 如何对指令建模？

由于面向 NLI 的指令和面向 LLM 的指令都与输入 X 或输出 Y 相关联，因此这些类型的指令不需要特定的系统设计来对其进行编码。面向 NLI 的指令可以由用于 NLI 任务的常规系统来处理，而面向 LLM 的指令大多被馈送到自回归 LLM。相反，面向人类的指令是最具挑战性的类型，因为它独立于任何已标记的实例。面向人类的指令建模策略包括：Semantic Parser [37, 60, 57]、Flatten-and-Concatenation [133, 140, 141] 、HyperNetwork [41, 45, 52, 156, 25, 48]、RLHF [125, 117, 3, 95]。

3.1.3 影响指令学习表现的因素

我们将影响指令学习性能的因素分为五个维度：模型、指令、示例、模型—指令对齐、数据集。

模型相关因素：（1）指令微调的 LLM > 普通 LLM。（2）指令使 LLM 变得更加安全、稳健和用户友好。（3）较大的 LLM 从指令学习中受益更多。

指令相关因素：（1）间隔几个 epochs 重写指令，直到它 work。（2）在训练和测试期间保持指令范式一致（例如抽象性）。（3）以不同的措辞和视角为一项任务设计多个指令。（4）提示词多样性简直让人精疲力尽？求助于 LLM 本身，尝试自动生成一些指令。（5）在大多数情况下，few-shot demonstrations 是有用的。

示例相关因素：我们将具体在3.2.2节中介绍。

模型—指令对齐相关因素：使用模型语言更好地设计你的指令（例如，遵循一定的相关目标）。

数据层面的相关因素：试着将 LLM 调整到更多样化的任务上。

3.1.4 挑战和未来方向

尽管指令学习具有上述提到的各种好处，但该领域仍然存在大量尚未探索的挑战。在本节中，我们列出了与遵循学习相关的几个挑战，这些挑战值得未来的研究调查。

（1）指令对齐的“代价”：除了追求性能之外，推理时的安全性也是指令微调的模型（或者说指令对齐）的一个重要方面。Ouyang et al. [95]用三个标准定义了“对齐”——有帮助、诚实和无害 (Helpful, Honest, and Harmless, HHH)，这已被近期的指令微调模型和数据集广泛考虑 [4, 158, 135, 78]。然而，对齐也会给指令调整模型带来“代价”。例如，Bekbayev et al. [5]发现，数据集指令中提供的很好对齐后的答案，可能会大大降低模型在各种任务基准上的性能。这意味着我们要仔细考虑在遵循指令的性能和安全性之间进行权衡。

（2）学习否定信息：否定是常见的语言属性，并且被发现对于各种 NLP 任务至关重要，例如 NLI[91, 53]。具体到指令遵循，否定表示上下文指令中任何需要避免的信息，包括否定要求（例如“避免使用停用词”）和否定示范（即一些错误的示例）。尽管人类可以从否定中学到很多东西 [30]，但现有的研究发现 LLM 常常无法遵循否定的指示；有些否定甚至会降低模型的性能 [62, 50, 88]。

（3）对抗性指令攻击：尽管大多数指令调整的语言模型可以很好地符合人类偏好并提供无害的响应，但最近的研究发现它们很容易受到攻击——可以通过使用简单的提示策略来操纵模型的响应。由于指令调整的 LLM 已应用于各种现实场景，例如网络代理和搜索引擎 [26, 147]，LLM 各代的安全性变得更加紧迫。仅仅进行偏好调整或内容过滤似乎是不够的，尤其是对于那些超强的 LLM 来说。因此，对于当前的指令调整模型来说，开发有效的防御方法是必要的。同时，进一步深入分析 LLM 的弱点也很重要，可能会提供更多关于防守的见解。

（4）指令学习的可解释性：正如我们在3.1.3中提到的，要实现更好的跨任务表现，关键因素之一是将以人为本的指令转换为面向 LLM 的指令，即让指令符合模型的偏好。许多以前的工作已经验证了在设计指令时迎合模型偏好的有效性，例如，在选择适当的指令时利用模型的困惑度 [38]。尽管这样做表现确实有所提高，但得到的指令始终违反人类的直觉，并显示出令人担忧的可靠性，例如一些语义不连贯、与任务无关、甚至具有误导性的指令 [55, 103]。这些结果证明了表现的提高和指令的人类可解释性之间的冲突，这是一个很难权衡的问题。

（5）学习遵循指令而不是仅仅生成标签预测 Y：为此，未来一个理想的方向是发展一种新的学习目标，以帮助 LLM 明确学习遵循指令，这可能会减轻对大规模标记实例的依赖。此外，还有一个更雄心勃勃、更具挑战性的想法是驱动系统实现学习指令，而不需要对任何特定任务的标记示例进行额外调整 [157]，这在某种程度上类似于基于语义解析器的范式。

（6）多语言的指令学习：直观上，指令学习是语言模型与语言或者说语种无关的能力，这意味着多语言的语言模型也有可能遵循不同语言的相同语义指令。例如， Kew et al. [54]发现使用三种以上语言调整的 LLM 表现出更强的指令学习能力，这意味着多语言指令调整的好处。不幸的是，当前大多数指令数据集和基础模型的开源指令都是以英语为中心的。因此，正如 Peng et al. [100]也提到的那样，高质量的多语言指令微调数据集（带有配对翻译）的发表对于未来的研究应该是有价值的。

3.2 上下文学习

ICL 中各研究领域的划分与概览图如图2所示，由于已经有非常全面完善的整理，本文中概览图便直接使用了 Dong et al. [29]一文中的。

3.2.1 训练阶段

在推理之前我们可以增加一个简单的训练阶段包括预训练（pretraining）和热身（warmup）。即怎样更好地激发模型 ICL 的能力。我们可以加一些预训练或者增量预训练，这个环节就是我们可以比如检索出和目标的 domain 相关度高的语料做 next token prediction 进一步预训练。然后我们还可以选择在预训练和推理之间加一点 warmup。指的是，用 ICL 模板做一下生成的训练，调整一下参数，可以缩小一下预训练和 ICL 之间的 gap。因为你预训练后实际上还是一个纯的自然语言的 next token prediction，它和按照 ICL 的模板去生成，之间还有一个 gap。

对于预训练，一个例子是 MEND [67]。整个过程 LLM 是冻结的，用一个 MEND 架构学习出 demonstration 的蒸馏部分，送到 Teacher LLM 来进行一个监督。就是说假设如果普通的预训练，用 ICL 模板这样的输入形式，训完后 LLM 对 demonstration 的表示的增量会怎么改变。这样就是我们只训一个很小的 MEND 去学这样一个东西。

对于 warmup，可以分成监督和自监督。监督的比如 metaICL 中取了 C 个 meta-training tasks，就是其他分类任务，不是他当前目标的这个任务。然后完全按照 ICL 推理时候的形式，取 k 个 demonstration 然后推 y_{k+1}，做这样的生成训练。推理或者说测试的时候是 unseen task。就是说这就是基于我们几乎没有目标任务的训练数据这样一个前提。然后它就发现，那我就用别的任务的数据，然后 ICL 的形式做一个 SFT，然后就发现能更好地激发 ICL 能力。这个直观来看，就是 warmup 让模型学会了知道去做这样一个 ICL 的类似归纳总结的事情，更好地知道并具备这样一个能力。自监督的比如 FLAN 按照一定规则自动化地去造一些 ICL 模板格式的监督任务，比如完形填空、下一句、下一个短语预测等等。

3.2.2 推理阶段

推理阶段包括对 demonstration 的处理、与一般的指令学习的结合、打分函数，这三个方向的研究。

demonstration 的处理包括 demonstration 的选择、demonstration 的重新格式化、demonstration 的顺序。

（1）demonstration 的选择：在偌大的数据集中，选择哪几条样本来做 demonstration 对表现最好，这是该方向所研究的问题。该方向的方法可以分为监督方法和无监督方法。无监督方法，如 KATE [70]，他们发现选择的 demonstration 与你待预测的输入相似度高会比较好。他们先对句子进行编码，然后对待预测的正式输入，通过 k 邻近来找到最近似的样本来做 demonstration。然后 knn 的过程中，相似度的打分，可以用 sentence-bert 也可以用 bert 在

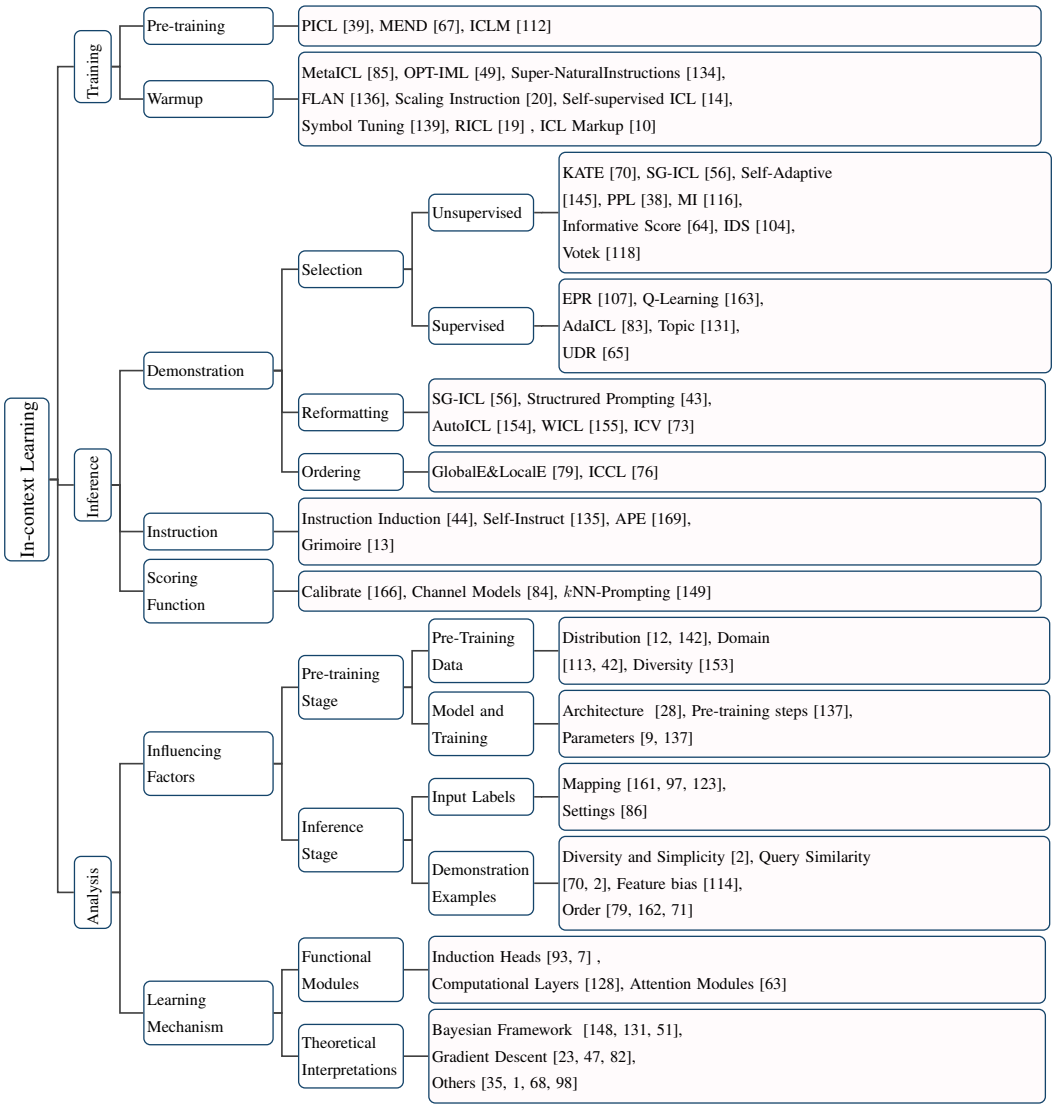

Figure 2: ICL 各研究领域概览图

[CLS] 处输出的 token。自监督方法，比如 EPR [107]，他们用 retriever，例如 sentence-bert，首先找出和输入 x 相似度最高的一些候选提示词，然后从目标的角度考虑，既然我们想要 x+ 合适的提示词来让 LLM 对标签 y 的预测最高，那么我们就看这些 x+ 提示词哪个能让 LLM 对标签 y 的预测最高，把这个作为分数，得到分数后，取 topk 标记为 1，取 bottomk 标记为 0，这样就得到了监督数据，然后就可以训练一个表示方式，结合 RAG 的方式索引、构建提示词并推理。

（2）demonstration 的重新格式化：即研究自动生成出我们想要的模板格式。例如 SG-ICL [56] 对待测样例，用 LLM 来生成 demonstration，再将 demonstration 拼接回去用于待测样例的预测。

（3）demonstration 的顺序：LLM 本质还是 next token prediction 的不断向后生成的模式，这导致它对 demonstration 的顺序实际上是很敏感的。例如 GlobalE & LocalE [79] 中，对 n 个 demonstration 构成的 ICL 模板，对比了 n! 个 demonstration 构成的排列的集合。他们发现，他们提出的两种熵都取得最大值时，即用作 demonstration 的样本的标签均匀分布的时候，预测的效果最好。

对于与一般指令学习的结合。当我们的目标很特别的时候，单纯的 ICL 无法实现，这时便不必拘泥于形式，将指令学习和 ICL 结合起来往往可能会得到一个想要的比较好的结果，毕竟，大模型本身还是对指令还是有着一定的理解能力。例如，一个正常的 ICL 过程，就是不断地 Input/Output 的模板格式，现在，你可能想知道给定 ICL 模板，隐含的任务或者说未知

指令是什么，那你就需要一个指令诱导的方式，让它回答未知指令是什么。如果不这样调整格式，按照大模型的特点，他就会自己继续向下生成 Input/Output 等内容。另外，思维链（chain of thought, CoT）也是一个例子，我们给问题和答案，这是 ICL，然后加入一些推导过程的思维链，这就是 ICL 与一般的指令学习的结合。

对于打分函数。即研究候选答案按照什么样的打分指标进行选择更好。最直接的方式是概率或者困惑度。Channel [84] 提出了一个很新颖的方法，它是一种类似贝叶斯估计的想法，对于 demonstration 和 y，判断在哪个 y 下，生成的 x 概率最高或者说困惑度最小。再比如 Calibrate [166] 中提出语料库中，例如，本来对下一个词是 positive 还是 negative 就有着一个天然的概率分布，或者说有一个天然的偏见（bias）。于是他们提出最后先输入一个空的 input，得到的各标签的生成概率即先验分布，就是这种天然的偏见，正式推理时除以掉这个概率即可。

3.2.3 理论分析

理论分析方向主要关注的是分析和 ICL 的表现强关联的影响因素，以及为什么 ICL work。影响因素，可以分成预训练和推理两个阶段分别探讨。

对于预训练阶段，有研究发现源域比语料库大小更有用，拼接多个语料库可能导致 ICL 能力的涌现；有研究分析了任务多样性的阈值，来使 LLM 在未见过的任务中展现 ICL 能力；有研究发现特定分布属性可能导致 ICL 能力涌现，例如突发性，某词的突然出现而不是随时间均匀分布；也有研究指出，模型参数够大或者预训练 steps 够多，就会涌现 ICL 能力，经验上来说 6B 一般是个起点，每个参数需要 1 到 10 个 tokens。

对于推理阶段，有研究证明了 input-label 的设置，比如配对格式以及输入的分布，显著影响 ICL 的 performance；有研究指出，准确的答案映射，影响很大；许多研究表明，Demonstration 的多样性、清晰度、以及顺序，有很明显的影响；也有研究指出，克服强大的先验偏差依然是个巨大的挑战。

对于学习机制的分析，可以分成对功能模块的分析以及对理论解释的分析。注意力模块一直是研究重点。有研究发现特定的注意力头，他们称作“induction head”归纳头，可以为 next token prediction 复制先前的 patterns；也有研究指出，demonstration 的 label 作为锚点，预测和分发了最终预测的关键信息。

理论上的解释，可以分成贝叶斯视角、梯度下降视角、以及解耦、算法学习、信息论等其他视角。梯度下降视角的一个简单例子如公式3 4。

$$
\begin{aligned}
\mathcal{F}(\mathrm{x}) &= (W_0 + \Delta W)\mathrm{x} \\
&= W_0\mathrm{x} + \Delta W\mathrm{x} \\
&= W_0\mathrm{x} + \sum_i \left(\mathrm{e}_i \otimes \mathrm{x}'_i\right)\mathrm{x} \\
&= W_0\mathrm{x} + \sum_i \mathrm{e}_i \left(\mathrm{x}'^T_i \mathrm{x}\right) \\
&= W_0\mathrm{x} + \text{LinearAttn}\left(E, \right)
\end{aligned}
\tag{3}
$$

$$
\begin{aligned}
\mathcal{F}_{\mathrm{ICL}}(\mathrm{q}) &= \mathrm{Attn}(V, K, \mathrm{q}) \\
&= W_V[X'; X]\ \mathrm{softmax}\left(\frac{(W_K[X'; X])^T \mathrm{q}}{\sqrt{d}}\right) \\
&\approx W_V[X'; X]\,(W_K[X'; X])^T \mathrm{q} \\
&= W_V X\,(W_K X)^T \mathrm{q} + W_V X'\,(W_K X')^T \mathrm{q} \\
&= \widetilde{\mathcal{F}}_{\mathrm{ICL}}(\mathrm{q}) \\
&= W_{\mathrm{ZSL}}\mathrm{q} + W_V X'\,(W_K X')^T \mathrm{q} \\
&= W_{\mathrm{ZSL}}\mathrm{q} + \mathrm{LinearAttn}\,(W_V X', W_K X', \mathrm{q}) \\
&= W_{\mathrm{ZSL}}\mathrm{q} + \sum_i W_V \mathrm{x}'_i \left((W_K \mathrm{x}'_i)^T \mathrm{q}\right) \\
&= W_{\mathrm{ZSL}}\mathrm{q} + \sum_i \left((W_V \mathrm{x}'_i) \otimes (W_K \mathrm{x}'_i)\right) \mathrm{q} \\
&= W_{\mathrm{ZSL}}\mathrm{q} + \Delta W_{\mathrm{ICL}}\mathrm{q} \\
&= (W_{\mathrm{ZSL}} + \Delta W_{\mathrm{ICL}})\mathrm{q}
\end{aligned}
\tag{4}
$$

他们计算了一个注意力头的参数 W 的增量改变，计算了 fine-tuning 会对模型的带来的改变量，然后通过推导可以发现，ICL 的增量和微调是类似的，因此即相当于一个隐式的梯度下降，让 ICL 不用训练，也能够发挥类似的作用。

3.2.4 ICL 的应用

由于用户友好的界面和轻量级的提示方法，ICL 在传统 NLP 任务上有着广泛的应用。近年来也衍生出了一些新兴的流行应用：

（1）数据工程：不像人工标注以及传统的有噪音的自动化标注，ICL 可以以更低的成本生成相对质量更高的数据，从而提高表现。

（2）模型增强：ICL 有着上下文灵活性有助于模型增强，比如可以结合应用于 RAG。

（3）知识更新：在预训练语料库中不存在的知识，大模型会不知道、无法回答、或者出现幻觉的回答，而 ICL 可以提供相关知识很好地解决这个问题。

（4）多模态：ICL 在 NLP 的巨大成功也鼓励了研究人员探索其在文本外各种模态的潜能，比如 vision、vision-language、speech 等。比如有研究采用 MAE（masked auto-encoders）进行图像块的填充，利用 ICL 在推理时生成一致的图像，展现了 ICL 强大的能力。再比如有研究者提出了 Prompt Diffusion model，是第一个具有 ICL 能力的 diffusion-based model，由额外的文本提示实现更精确的图像生成。

3.2.5 ICL 的局限性和未来挑战

（1）效率和可扩展性：效率上，随着演示数量的增加，计算成本会更高。可扩展性上，由于 LLM 的最大输入长度，可学习样本会比较少。（2）泛化问题，ICL 依赖于高质量的 demonstration，而高质量的 demonstration 是稀缺的，这种稀缺性对 ICL 的泛化能力提出了挑战。（3）对于长文本，实验表明，增加 demonstration 的数量并不一定会提高表现，甚至可能有害。与此同时，关于其导致表现下降的原因也还需要进一步的研究与调查。

3.3 自然语言推理

本节中，主要总结自然语言推理中的一些方法。截至目前，关于自然语言推理，研究者们已经开发了许多方法。这些方法的范围早期的符号和统计方法到最近应用深度学习和神经网络的方法。符号方法使用逻辑形式和过程来进行推理。统计方法则是从 20 世纪 90 年代中期到 2010 年代初，在 NLP 领域占据主导地位。这些早期的统计方法依赖于工程特征来根据各种早期 NLP 基准的训练数据来训练各种类型的统计模型。直到近十年神经网络方法成为主流，神经网络方法从早期的统计方法发展而来，但使用各种神经网络架构来发现数据中

的有用特征，而不是手动指定所有特征。最近的基准测试数据量越来越大，使得训练更大、更深的神经模型成为可能。如今，这些方法几乎名列此处调查的所有 NLI 基准排行榜的首位。传统的词嵌入模型中，例如 word2vec [21] 或 GloVe [101]，嵌入向量与上下文无关。近年来的 Transformers 类模型，比如 BERT，应用了双向的编码器。些模型根据单词出现的上下文为单词提供不同的嵌入向量。这些预先训练的单词表示可以用作特征或针对下游任务进行微调。以斯坦福语义推理为例，它们从 LSTM/CNN 一直衍生到现在的 BERT，测试准确率从最初的 77.6% 一直到如今的 93.1%，，我们整理了近七年使用了注意力机制的一些文章 [74, 36, 124, 96, 59, 36, 124, 122, 96, 59, 164, 75, 165, 102, 121, 130]

3.4 命名实体识别

本节中，主要总结命名实体识别中的一些方法、面对的挑战和未来的方向。截至目前，关于命名实体识别，研究者们已经开发了许多方法。这些方法的范围早期的符号和统计方法到最近应用深度学习和神经网络的方法。最常用的机器学习方法是隐马尔可夫模型 (HMM) [168]、条件随机场 (CRF) [99] 和支持向量机 (SVM) [115]。深度学习中，最常用的方法是循环神经网络如 BiLSTM，然后 CNN 以及二者的结合。通常使用的单词表示是 One-Hot 编码、Count Vectorizer、TF-IDF、Word2vec 方法（例如 Skip gram）以及第二连续词袋 (CBOW)。我们整理了近些年命名实体识别中的一些方法 [33, 6, 92, 152, 119, 34, 17, 31]，他们的测试 F1 分数从 64% 一直到约 91%。

目前，NER 领域面对的挑战有：

（1）数据注释：有监督的 NER 系统，特别是基于 DL 的 NER 系统，需要大量带注释的训练数据。注释数据仍然费时又费钱。由于需要领域专家来完成注释活动，这对于许多语言和需要更多资源的特定领域来说是一个重大障碍。

（2）复杂的生物医学文本：在生物医学文献中，冗长而复杂的句子很常见。有时，两个复杂的实体有时可能出现在两个单独的子句中。准确地识别和分类复杂的生物医学句子中的相关生物医学关系是生物医学关系分类的一个重大困难。

（3）语言中的歧义性：自然语言单词可以有多种解释，因此很难确定一个术语是否对应于指定事物或其他事物。例如句子 “He saw a bat”，这个简单的句子可以有四种不同的含义，因为单词 “saw” 和 “bat”。锯子可以有单词 see 和 Cutting 的过去式两种含义，蝙蝠可以有两种含义：鸟或运动器材。因此，该句子可以用四种不同的方式解释。

（4）非正式文本：由于其简洁性和噪音，非正式文本（例如评论、推文和用户论坛）上的 NER 比正式文本提出了更多的挑战。NER 系统必须在各种应用程序设置中操作用户生成的文本，包括电子商务和银行中的客户服务。它们的句子结构较弱，因此很难使用通用特征来定位命名实体。

（5）多语言 NER：全球互联网用户的增加，将多种文化、人民和语言聚集在一起，使其在语言上变得多样化。然而，必须进行更多的研究来解决网络材料日益增长的语言多样性问题。自然语言处理技术仅限于少数语言，通常仅限于英语，这跟不上互联网的快速变化。因此，当前基于多语言文本的 IR 系统仅限于利用基于词频的技术和表面形式的基本处理阶段。

（6）域适应：跨域 NER 是一个困难但可行的问题。跨领域，实体提及可能会有很大差异。例如，在植物学中，“橙色” 既可以指水果，也可以指颜色。同样，不同类型的实体与其他学科中其他类型的实体的相似性可能有所不同。例如，疾病在医学和生物化学方面的治疗方式有所不同。

（7）命名实体链接：NER 的挑战之一是实体链接。实体可能具有不同的含义，并且在文档中的引用方式也不同。例如，“巴塞罗那” 在同一份文件中可以指一支足球队和一座城市。

（8）实体共指解析：确定文本或对话中提到的实体是否引用同一现实世界实体称为实体共指解析。

（9）处理嘈杂和拼写错误的文本：大多数文档都存在清晰度问题，例如，由于某些 OCR 问题拼写错误，从 OCR 读取的扫描文档不清晰，常见的看起来相似的字母被错误检测并导致拼写错误。人为错误也可能导致文本中出现拼写错误。

未来的方向有：

名称	demonstration 格式举例
origin	Input:This quiet , introspective and entertaining independent is worth seeking . Output:Positive
origin_with_space	Input: This quiet , introspective and entertaining independent is worth seeking . Output: Positive
origin_with_newline	Input: This quiet , introspective and entertaining independent is worth seeking . Output: Positive
sentiment	Sentence: This quiet , introspective and entertaining independent is worth seeking . Sentiment: Positive
sentiment_channel	Sentiment: Positive Sentence: This quiet , introspective and entertaining independent is worth seeking .

Table 2: task1/task2 实验尝试的 demonstration 模板格式举例

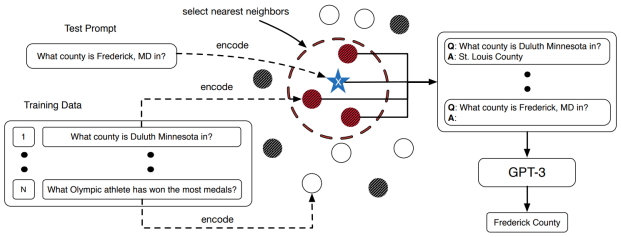

Figure 3: KATE 方法概述图

(1) 由于每种模型或方法都有其方法论上的优点和缺点，因此组合各种模型或方法可能会产生卓越的成果。在关系分类中，一些混合模型已经证明了将 CNN 与 RNN 合并的可能性。由于 RNN 无法同时识别多个句子，因此不可能获得完整的信息。

(2) 创建能够识别未经明确培训的语言中的命名项的模型是零射跨语言 NER 的目标。该方法可以大大降低低资源语言对标注数据的需求，并能有效地在训练数据较少的语言中使用 NER。

(3) 新模型和技术可以与基于神经网络的方法相结合，以进一步提高效率。深度上下文化单词表示，如 ELMo 和 BERT，网络拓扑可以很容易地针对分类任务进行修改，例如序列标记活动（例如命名实体识别）、生产任务（例如如抽象摘要、垃圾邮件检测和其他类型的任务。许多其他迁移学习方法也可以通过集成现有模型或独立使用来获得更好的结果。

(4) 迁移学习：当我们拥有相对较小的数据集时，迁移学习是首选。迁移学习旨在提高模型的泛化性，并最大限度地减少目标任务的训练标记数据。

4 实验方法

4.1 影评情感分类

影评情感分类是 NLP Beginner 的 task1/task2，它旨在给定一个 phrase，判断句子的情感极性，包括：{negative, somewhat negative, neutral, somewhat positive, positve} 五个类别。在 task1 中我们首先完成了机器学习方法的相关对比实验。然后在这个过程中，最后的处理过程和采用的方法如下：首先，对于数据清洗和预处理，这个任务中比较坑的是，训练集有两个样本 phrase 只有空白符，测试集中有个 phrase 居然是 None，清洗之后，我们做了一些预处理，包括删除了单词字符 (a-z,A-Z,0-9) 和空白符外的所有字符，删除了全部长度 <=2 的停

名称	demonstration 格式举例
NLI	Premise: A soccer game with multiple males playing. Hypothesis: Some men are playing a sport. Relationship: Entailment
NLI_channel	Relationship: Entailment Premise: A soccer game with multiple males playing. Hypothesis: Some men are playing a sport.

Table 3: task3 实验尝试的 demonstration 模板格式举例

部分	内容
prefix	Now I will ask you to answer the relationship between the input premise and hypothesis. Please answer my question based on the following samples of premise, hypothesis and relationship. In this question the relationship between the premise and hypothesis includes three categories: Contradiction, Neutral, Entailment. Contradiction means that the premise and hypothesis are contradictory, that is, the premise can clearly infer that the hypothesis is not true. Neutral means that there is no clear direct connection between the premise and the hypothesis, that is, the premise can neither clearly infer that the hypothesis is true nor clearly infer that the hypothesis is not true. Entailment means that the hypothesis is included in the premise, that is, the premise can clearly infer that the hypothesis is true. What you need to do is to answer the relationship between the input premise and hypothesis
demonstrations	`[demonstrations]`
infix	The following is the formal input. Please imitate the above samples and answer my question as required:
query	`[query]`

Table 4: task3 实验尝试加入一般指令后的 ICL 模板格式

用词（stop words），对所有单词做了词形还原（lemmatize）。然后最终采用了 TFIDF 来表示特征，并用 LGBM 进行分类，提交 Kaggle。然后在 task2 中，我们复现了 TextCNN [106]，分别训练了 Glove+TextCNN、Glove+BiLSTM、微调 RoBERTa 完成了该分类任务，并提交 Kaggle。然后我们还尝试使用 ICL 方法，首先尝试调整模板，如表2所示。尝试的原始格式是 Input/Output，然后分别尝试了在冒号后加空格和直接加换行。由于发现加空格表现最好，保留这个格式，又进一步尝试了将 Input/Output 改为 Sentence/Sentiment 的形式。对于 demonstration 的数量，我们进行了 5-shot 和 10-shot 的对比。对于 demonstration 的标签分布问题，根据 GlobalE & LocalE [79] 一文给出的结论，我们始终选择示例各标签均匀分布。然后我们还复现了 Channel [84], Calibrate [166], KATE [70]（图3）并进行了对比分析。关于选取的 ICL 方法的详细信息，我们已在3.2.2中介绍。对于推理使用的大模型，统一使用的是 llama-3.2-1b-instruct 进行对比。

4.2 自然语言推理

斯坦福自然语言推理是 NLP Beginner 的 task3，它旨在给定前提 premise（sentence1）和假说 hypothesis（sentence2），然后判断二者的推理关系，包括：{contradict, neutral, positive} 三个类别。在这个任务中，我们首先完成了复现 ESIM [15]（图4，忽略 Tree-LTSM 的部分）的基本要求，然后复现了 Self-Explain RoBERTa [121]（图5）进行对比。接下来尝试使用 ICL 方法，由于有了上个任务的经验，我们这次直接采用清晰的提示词以及冒号后有空格的格式，如表3所示。

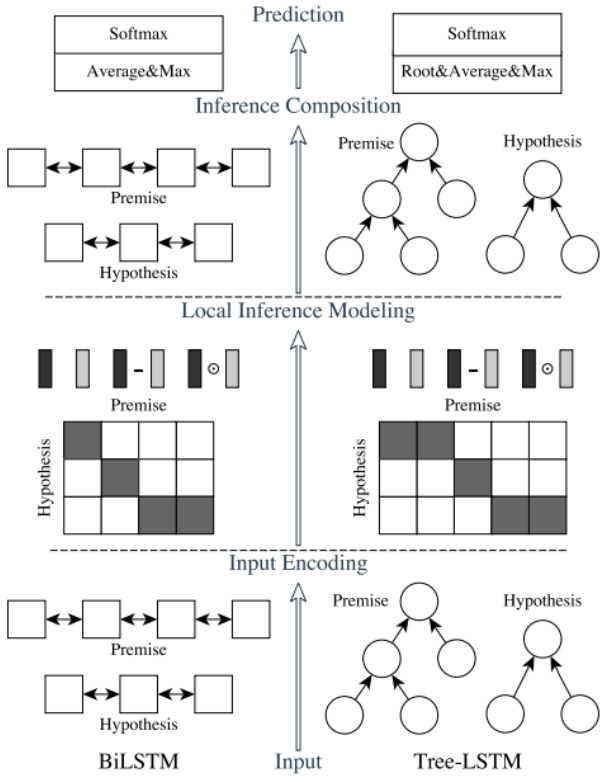

Figure 4: ESIM 模型结构

由于相比上一个任务，这个任务的测试集相对小许多，推理时间更短。我们还进一步尝试加入一些一般形式的指令，并对这些指令进行了消融实验，完整的指令格式如表4所示。

然后在这个任务中，我们也对比了 3-shot 和 6-shot。并且也同样复现了 Channel [84], Calibrate [166], KATE [70]（图3）并进行了对比分析。

4.3 命名实体识别

命名实体识别是 NLP Beginner 的 task4，它旨在为给定文本已经分好的各词识别每词对应的具有特定意义的实体类别，在这个任务中，使用的数据集来自 CONLL 2023，9 个类别，包括 {B-LOC, B-MISC, B-ORG, B-PER, I-LOC, I-MISC, I-ORG, I-PER, O}。任务中，我们首先复现了 LSTM+CRF [80]，然后在此基础上将 LSTM 改用 bert-based-cased，因为某些首字母大写的词往往有特定意义，会对实体类别产生影响，因此采用 cased 版本。

4.4 中文唐诗生成

在该任务中，有一个很小的唐诗数据集作为训练集。我们首先按照要求训练 LSTM 生成唐诗。然后也直接使用了已经预训练好的 gpt2-chinese-poem 做生成来对比。生成过程中，对 gpt2 也尝试和 generate 函数结合使用。尝试随机生成以及限定格式生成，格式包括每一句字数和句数，以及藏头诗。由于不是本文的主要实验，结果展示在附录A中。

5 实验结果与分析

首先，关于合适的 ICL 模板的探索，在 task1/2 中，我们尝试了不同的 demonstration 的提示格式。然后在 task3 中我们进一步探索了结合一般指令构成模板。具体设置在第4.14.2节中进行了详细的介绍。对于 demonstration 的提示格式的探索，实验结果在表5中，通过实验我

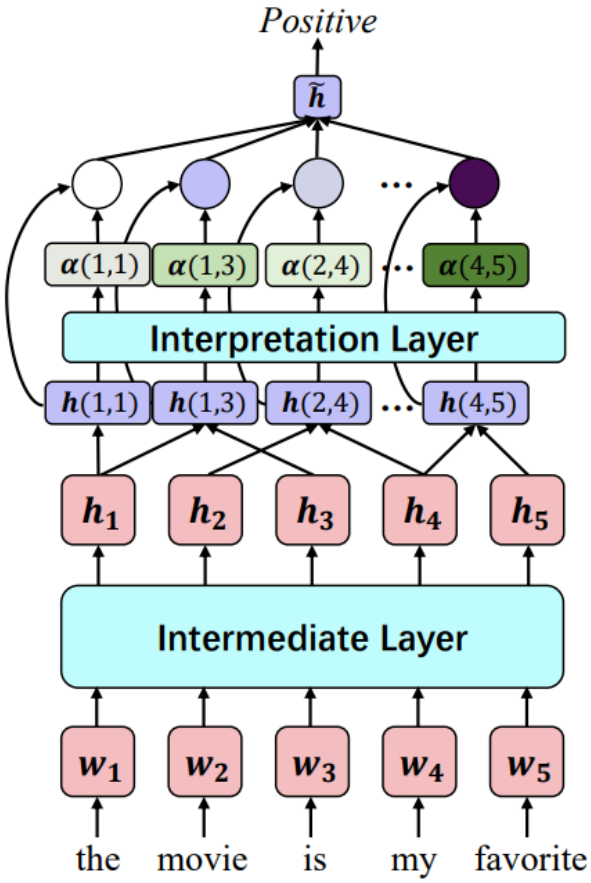

Figure 5: Self-Explain RoBERTa 模型结构

格式名称	acc(%)	F1(%)
origin	4.60	1.84
origin_with_space	29.04	24.26
origin_with_newline	19.98	19.22
sentiment	**44.08**	**31.43**

Table 5: task1/2 实验尝试不同 demonstration 格式的实验结果（格式名称对应形式的例子见表2）

Submission and Description	Private Score	Public Score	Selected
submission_KATE-llama-3.2-1b-instruct-sentiment.csv Complete (after deadline) · 2d ago	0.41501	0.41501	
submission_ICL-llama-3.2-1b-instruct-sentiment.csv Complete (after deadline) · 4d ago	0.46169	0.46169	
submission_lgbm.csv Complete (after deadline) · 12d ago	0.51789	0.51789	
submission_roberta.csv Complete (after deadline) · 13d ago	0.69869	0.69869	
submission_lstm.csv Complete (after deadline) · 13d ago	0.66196	0.66196	
submission_cnn.csv Complete (after deadline) · 13d ago	0.64567	0.64567	

Figure 6: task1/2 Kaggle 测试提交结果

前缀/中缀形式	acc(%)	F1(%)
不增加一般指令的原始形式	43.99	33.86
增加一般指令后的完整格式	38.26	29.51
将前缀以句号结尾的倒数第 2 到 4 句精简（替换）为", namely, whether the premise can infer the hypothesis."	38.26	29.51
在表格上一行精简的基础上去掉前缀最后一句	40.00	31.51
在表格上一行的基础上去除中缀	41.97	31.24
去除前缀保留中缀（但从人类语言理解上来看有误）	**45.50**	**36.17**
去除前缀并将中缀修改为"The following is the formal input. Please imitate the above samples and answer my question:"	45.12	35.89
在 task1/2 中尝试 task3 表现最好的仅保留中缀的形式	36.14(↓7.94)	27.61(↓3.82)

Table 6: task3 实验尝试加入一般指令并消融的实验结果（完整格式见表4）

示例数 (shot)	acc(%)	F1(%)
5-shot(task1/2)	**44.08**	**31.43**
10-shot(task1/2)	30.19	29.62
3-shot(task3)	**43.99**	**33.86**
6-shot(task3)	39.25	32.57

Table 7: task1/2、task3 不同示例个数实验结果

类别	任务	方法	acc(%)	F1(%)
训练方法	task1/2	$\text{LGBM}_{\text{TFIDF}}$	50.99	44.28
		$\text{TextCNN}_{\text{GLOVE}}$ [106]	67.99	57.46
		$\text{LSTM}_{\text{GLOVE}}$	68.80	57.91
		$\text{RoBERTa}_{\text{base}}$	**70.31**	**62.29**
	task3	ESIMESIM [15]	82.82	82.82
		Self Explain $\text{RoBERTa}_{\text{base}}$ [121]	**91.08**	**91.04**
	task4	LSTM+CRF [80]	95.96	83.14
		$\text{BERT}_{\text{base-cased}}$ + CRF	**96.69**	**85.74**
ICL	task1/2	ICL	**44.08**	**31.43**
		Channel-ICL [84]	25.74	22.07
		Calibrate [166]	19.72	17.40
		KATE [70]	31.70	19.15
	task3	ICL	43.99	33.86
		Channel-ICL [84]	**46.76**	**46.57**
		Calibrate [166]	37.19	32.38
		KATE [70]	45.81	38.67

Table 8: NLP Beginner 最终实验结果准确率与 F1（task1/2 取本地验证结果）

们证明了 **ICL 对模板格式高度敏感**，原始的 Input/Output 并且冒号后无空格的形式有着极低的准确率，和实验中最好的模板格式差别高达十倍。实验表明冒号后加一个空格要明显更好，并且比直接换行也要好许多。一个在简洁的前提下尽可能清晰确切的提示，也是提高表现的一个关键，比如在情感分类中将"Input/Output" 更改为"Sentence/Sentiment" 可以明显提高表现。然后 task3 中我们结合一般指令构成模板的实验结果在表6中。实验证明，**ICL 对模板长度是高度敏感的，即便解释清晰，但过长的模板反而会导致性能下滑**。如何用最简洁最清晰最确切地方式为大语言模型提供最佳的提示，也是一个巨大的挑战，毫无疑问，这会明显影响表现。实验结果表明，将前缀指令全部去除可以取得最佳表现。另外，令我们感到惊恐地是，意外失误仅保留中缀取得了最好的表现，中缀内容是"The following is the formal input. Please imitate the above samples and answer my question as required:"（见表4），可是我们已经去掉了前缀中的 require 了，表现的提升令我们感到恐惧，我们尝试将其调整为人类语言上的正确形式6，但结果却出现了一点点下滑，不过仅加入中缀还是带来了表现的提升，因此我们也尝试在 task2 中的模板加入中缀，但结果却出现了一定的下滑6，这让我们很费解，加入中缀来明确区分 demonstration 和正式输入或许某些时候确实可能会奏效，但似乎并不总是有效，而且实验中的意外让人无法理解，因此**在最终实验结果8的对比中我们选择不对比这个意外**。另外，我们在 task1/2、task3 中还进行了 few-shot 示例数量的对比，实验表明，示例数量尽可能少反而往往会有更好的结果，这也再次证明了 ICL 对模板长度的高度敏感，过长的模板反而会导致表现下降。

NLP Beginner 的各训练方法及各 ICL 方法最终实验结果对比在表8。对于 task1 中原要求的机器学习对比实验，不在这里具体展开赘述。对于 task1/2 在 Kaggle 的测试结果，在图6中。众所周知的是，神经网络总是容易过拟合，而这尤其对竞赛问题的影响十分大，测试结果往往会明显低于验证结果。需要注意的是，为了保证泛化能力，task1/2 中，为了保证更好的泛化能力，我们尽可能不训练过多的 epochs 来通过在极值点附近的抖动追求验证的极限，而是保证刚收敛即可，并且我们的 batch_size 也尝试设置相对较小的合适值，并且我们还在一定程度内尽可能加大 weight_decay 和 dropout_rate。因此，值得高兴的是，例如，task2 RoBERTa 本地验证准确率 70.31%，kaggle 提交结果 69.87%，这是非常接近的，但即便如此，依然是下降的。而相比之下，**ICL 表现出了强大的泛化能力**，因为其本质上总体来说相比于训练的方法是不存在过拟合的影响的，甚至在这个任务中，在 kaggle 的测试提交结果出现了明显的上升。但遗憾的是，我们的实验也证明了，**在具有一定的难度的特定分类任务中，未经过指令微调的大模型仅通过 ICL 方法表现是极其有限的**，远不如过拟合的分类器有效。

对于复现的一些尝试在推理阶段，通过改进 demonstration 选择和打分方式的一些 ICL 方法。在表8中可以看到，task3 中 Channel-ICL 取得了最佳的表现，KATE 的表现也非常不错，当对原始的 ICL 表现不够满意时，Channel-ICL 和 KATE 是值得一试的，尽管可能并非总是表现良好，例如在 task1/2 中。Calibrate 的表现并不好，去除自然语言对各标签的天然偏见并没有带来更好的表现，我们推测，**生成模型中的自然语言偏见某些情形下可能是必要的**。毕竟我们注释的标签可能本身就带有标签，而这往往也可能是整个人类社会本身带有的一些偏见，人类的客观往往可能并不是真正的客观，但有时我们或许恰恰应当需要这种有偏见的客观。但是，或许在 ScienceQA 中，Clibrate 能够取得明显更好的表现，这需要进一步实验来证明。此外，我们还尝试了将 Channel 与 Calibrate 方法结合，但是相比于单独的 Channel，表现没有任何差别，因此没有记录在表中，这或许证明了 **Channel 方法本身就是几乎不受偏见影响的，与偏见几乎无关**。事实上，我们认为 task1/2 中的偏见是极其明显的，仅是人类观察就能明显感受到，这或许也是为什么 Channel 方法在 task1/2 中表现并不是很好。或许相比于原生的 ICL，Channel 在大部分情况下是一个非常有效的方法，但这同样需要更多的调查。

6 总结

我们的期末作业报告，首先完成了对指令学习（Instructio Learning）、上下文学习（ICL）、自然语言推理（NLI）、命名实体识别（NER）的综述。对于指令学习，我们总结了什么是指令、怎样构建指令、影响指令学习的因素，并基于总结指出了挑战和未来方向。对于上下文学习，我们从训练阶段、推理阶段、理论分析角度总结了 ICL 目前的各研究领域，然后也概述了 ICL 目前的一些新兴应用，并指出了 ICL 目前的局限性和未来的挑战。对于自然语言推理和命名实体识别，我们主要总结了其从统计学习方法为主流的时期至神经网络为主流的时期所应用的一些方法。接着，我们完成了本文的实验报告，在完成 NLP Beginner 原任务以及改用 BERT 类模型的基础上，进一步对 task1/2、task3 尝试了一些 ICL 的方法，并得到了摘要中给出的一些结论。ICL 展现出了神经网络训练方法所不具有的许多优点，但同样也有着明显的弊端，但通过自然语言提示来解决问题的方式毫无疑问是更加接近真实世界的。

我们对 ICL 乃至其他指令学习相关的方法保持乐观的态度，并希望未来这些方法能够得到更好的发展与应用。

References

[1] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=0g0X4H8yN4I.

[2] Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. How do in-context examples affect compositional generalization? In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11027–11052. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.618. URL https://doi.org/10.18653/v1/2023.acl-long.618.

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862, 2022. doi: 10.48550/ARXIV.2204.05862. URL https://doi.org/10.48550/arXiv.2204.05862.

[4] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073, 2022. doi: 10.48550/ARXIV.2212.08073. URL https://doi.org/10.48550/arXiv.2212.08073.

[5] Aibek Bekbayev, Sungbae Chun, Yerzat Dulat, and James Yamazaki. The poison of alignment. *ArXiv preprint*, abs/2308.13449, 2023. URL https://arxiv.org/abs/2308.13449.

[6] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[7] Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. Birth of a transformer: A memory viewpoint. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/0561738a239a995c8cd2ef0e50cfa4fd-Abstract-Conference.html.

[8] 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. Language models are few-shot

learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[9] 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. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[10] Marc-Etienne Brunet, Ashton Anderson, and Richard S. Zemel. ICL markup: Structuring in-context learning using soft-token tags. *CoRR*, abs/2312.07405, 2023. doi: 10.48550/ARXIV.2312.07405. URL https://doi.org/10.48550/arXiv.2312.07405.

[11] Thomas P Carpenter, Elizabeth Fennema, and Megan L Franke. Cognitively guided instruction: A knowledge base for reform in primary mathematics instruction. *The elementary school journal*, 1996.

[12] Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya K. Singh, Pierre H. Richemond, James L. McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/77c6ccacfd9962e2307fc64680fc5ace-Abstract-Conference.html.

[13] Ding Chen, Shichao Song, Qingchen Yu, Zhiyu Li, Wenjin Wang, Feiyu Xiong, and Bo Tang. Grimoire is all you need for enhancing large language models. *CoRR*, abs/2401.03385, 2024. doi: 10.48550/ARXIV.2401.03385. URL https://doi.org/10.48550/arXiv.2401.03385.

[14] Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. Improving in-context few-shot learning via self-supervised training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3558–3573, Seattle, United States, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.naacl-main.260.

[15] Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*, 2016.

[16] Nancy Chinchor and Patricia Robinson. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, volume 29, pages 1–21, 1997.

[17] Minsoo Cho, Jihwan Ha, Chihyun Park, and Sanghyun Park. Combinatorial feature embedding based on cnn and lstm for biomedical named entity recognition. *Journal of biomedical informatics*, 103:103381, 2020.

[18] 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. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113, 2023. URL http://jmlr.org/papers/v24/22-1144.html.

[19] Timothy Chu, Zhao Song, and Chiwun Yang. Fine-tune language models to approximate unbiased in-context learning. *CoRR*, abs/2310.03331, 2023. doi: 10.48550/ARXIV.2310.03331. URL https://doi.org/10.48550/arXiv.2310.03331.

[20] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

[21] Kenneth Ward Church. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017.

[22] Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.161. URL https://aclanthology.org/2021.findings-acl.161.

[23] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4005–4019. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.247. URL https://doi.org/10.18653/v1/2023.findings-acl.247.

[24] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103, 2015.

[25] Budhaditya Deb, Ahmed Hassan Awadallah, and Guoqing Zheng. Boosting natural language generation from instructions with meta-learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6792–6808, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.456.

[26] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samual Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/5950bf290a1570ea401bf98882128160-Abstract-Datasets_and_Benchmarks.html.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[28] Nan Ding, Tomer Levinboim, Jialin Wu, Sebastian Goodman, and Radu Soricut. CausalLM is not optimal for in-context learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=guRNebwZBb.

[29] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, et al. A Survey on In-Context Learning. *ArXiv preprint*, abs/2301.00234, 2023. URL https://arxiv.org/abs/2301.00234.

[30] Carolin Dudschig and Barbara Kaup. How does "not left" become "right" ? electrophysiological evidence for a dynamic conflict-bound negation processing account. *Journal of Experimental Psychology: Human Perception and Performance*, 44(5):716, 2018.

[31] Shengyu Fan, Hui Yu, Xiaoya Cai, Yanfang Geng, Guangzhen Li, Weizhi Xu, Xia Wang, and Yaping Yang. Multi-attention deep neural network fusing character and word embedding for clinical and biomedical concept extraction. *Information Sciences*, 608:778–793, 2022.

[32] Elizabeth Fennema, Thomas P Carpenter, Megan L Franke, Linda Levi, Victoria R Jacobs, and Susan B Empson. A longitudinal study of learning to use children's thinking in mathematics instruction. *Journal for research in mathematics education*, 1996.

[33] J Fries, S Wu, A Ratner, and C Ré. A generative model for biomedical named entity recognition without labeled data. *arXiv preprint arXiv:1704.06360*, 2017.

[34] Sudhakaran Gajendran, D Manjula, and Vijayan Sugumaran. Character level and word level embedding with bidirectional lstm–dynamic recurrent neural network for biomedical named entity recognition from literature. *Journal of Biomedical Informatics*, 112:103609, 2020.

[35] Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/c529dba08a146ea8d6cf715ae8930cbe-Abstract-Conference.html.

[36] Reza Ghaeini, Sadid A Hasan, Vivek Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Yuan Ling, Aaditya Prakash, Xiaoli Z Fern, and Oladimeji Farri. Dr-bilstm: Dependent reading bidirectional lstm for natural language inference. *arXiv preprint arXiv:1802.05577*, 2018.

[37] Dan Goldwasser and Dan Roth. Learning from natural instructions. In Toby Walsh, editor, *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 1794–1800. IJCAI/AAAI, 2011. doi: 10.5591/978-1-57735-516-8/IJCAI11-301. URL https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-301.

[38] Hila Gonen, Srini Iyer, Terra Blevins, Noah A. Smith, and Luke Zettlemoyer. Demystifying prompts in language models via perplexity estimation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10136–10148. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.679. URL https://doi.org/10.18653/v1/2023.findings-emnlp.679.

[39] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Pre-training to learn in context. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4849–4870. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.267. URL https://doi.org/10.18653/v1/2023.acl-long.267.

[40] Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey Bigham. InstructDial: Improving zero and few-shot generalization in dialogue through instruction tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 505–525, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.33.

[41] David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=rkpACe1lx.

[42] Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. Understanding in-context learning via supportive pretraining data. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12660–12673. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.708. URL https://doi.org/10.18653/v1/2023.acl-long.708.

[43] Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. Structured prompting: Scaling in-context learning to 1,000 examples. *ArXiv preprint*, abs/2212.06713, 2022. URL https://arxiv.org/abs/2212.06713.

[44] Or Honovich, Uri Shaham, Samuel R. Bowman, and Omer Levy. Instruction induction: From few examples to natural language task descriptions. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1935–1952. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.108. URL https://doi.org/10.18653/v1/2023.acl-long.108.

[45] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 2019. URL http://proceedings.mlr.press/v97/houlsby19a.html.

[46] Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A. Smith, and Mari Ostendorf. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2627–2643, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.193.

[47] Kazuki Irie, Róbert Csordás, and Jürgen Schmidhuber. The dual form of neural networks revisited: Connecting test time predictions to training patterns via spotlights of attention. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 9639–9659. PMLR, 2022. URL https://proceedings.mlr.press/v162/irie22a.html.

[48] Hamish Ivison, Akshita Bhagia, Yizhong Wang, Hannaneh Hajishirzi, and Matthew E. Peters. HINT: hypernetwork instruction tuning for efficient zero- and few-shot generalisation. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11272–11288. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.631. URL https://doi.org/10.18653/v1/2023.acl-long.631.

[49] 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. Opt-iml: Scaling language model instruction meta learning through the lens of generalization, 2022. URL https://arxiv.org/abs/2212.12017.

[50] Joel Jang, Seonghyeon Ye, and Minjoon Seo. Can large language models truly understand prompts? A case study with negated prompts. In Alon Albalak, Chunting Zhou, Colin Raffel, Deepak Ramachandran, Sebastian Ruder, and Xuezhe Ma, editors, *Transfer Learning for Natural Language Processing Workshop, 03 December 2022, New Orleans, Louisiana, USA*,

volume 203 of *Proceedings of Machine Learning Research*, pages 52–62. PMLR, 2022. URL https://proceedings.mlr.press/v203/jang23a.html.

[51] Hui Jiang. A latent space theory for emergent abilities in large language models. *CoRR*, abs/2304.09960, 2023. doi: 10.48550/ARXIV.2304.09960. URL https://doi.org/10.48550/arXiv.2304.09960.

[52] Tian Jin, Zhun Liu, Shengjia Yan, Alexandre Eichenberger, and Louis-Philippe Morency. Language to network: Conditional parameter adaptation with natural language descriptions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6994–7007, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.625. URL https://aclanthology.org/2020.acl-main.625.

[53] Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.698. URL https://aclanthology.org/2020.acl-main.698.

[54] Tannon Kew, Florian Schottmann, and Rico Sennrich. Turning english-centric llms into polyglots: How much multilinguality is needed? *ArXiv preprint*, abs/2312.12683, 2023. URL https://arxiv.org/abs/2312.12683.

[55] Daniel Khashabi, Xinxi Lyu, Sewon Min, Lianhui Qin, Kyle Richardson, Sean Welleck, Hannaneh Hajishirzi, Tushar Khot, Ashish Sabharwal, Sameer Singh, and Yejin Choi. Prompt waywardness: The curious case of discretized interpretation of continuous prompts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3631–3643, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.266. URL https://aclanthology.org/2022.naacl-main.266.

[56] Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator. *ArXiv preprint*, abs/2206.08082, 2022. URL https://arxiv.org/abs/2206.08082.

[57] Joohyun Kim and Raymond Mooney. Unsupervised PCFG induction for grounded language learning with highly ambiguous supervision. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 433–444, Jeju Island, Korea, 2012. Association for Computational Linguistics. URL https://aclanthology.org/D12-1040.

[58] Juae Kim, Yejin Kim, and Sangwoo Kang. Weakly labeled data augmentation for social media named entity recognition. *Expert Systems with Applications*, 209:118217, 2022.

[59] Seonhoon Kim, Inho Kang, and Nojun Kwak. Semantic sentence matching with densely-connected recurrent and co-attentive information. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6586–6593, 2019.

[60] Gregory Kuhlmann, Peter Stone, Raymond Mooney, and Jude Shavlik. Guiding a reinforcement learner with natural language advice: Initial results in robocup soccer. In *The AAAI-2004 workshop on supervisory control of learning and adaptive systems*. San Jose, CA, 2004.

[61] Bangzheng Li, Wenpeng Yin, and Muhao Chen. Ultra-fine entity typing with indirect supervision from natural language inference. *Transactions of the Association for Computational Linguistics*, 10:607–622, 2022. doi: 10.1162/tacl_a_00479. URL https://aclanthology.org/2022.tacl-1.35.

[62] Judith Yue Li, Aren Jansen, Qingqing Huang, Ravi Ganti, Joonseok Lee, and Dima Kuzmin. Maqa: A multimodal qa benchmark for negation. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022.

[63] Shuai Li, Zhao Song, Yu Xia, Tong Yu, and Tianyi Zhou. The closeness of in-context learning and weight shifting for softmax regression. *CoRR*, abs/2304.13276, 2023. doi: 10.48550/arXiv.2304.13276. URL https://doi.org/10.48550/arXiv.2304.13276.

[64] Xiaonan Li and Xipeng Qiu. Finding supporting examples for in-context learning. *CoRR*, abs/2302.13539, 2023. doi: 10.48550/ARXIV.2302.13539. URL https://doi.org/10.48550/arXiv.2302.13539.

[65] Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. Unified demonstration retriever for in-context learning. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4644–4668. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.256. URL https://doi.org/10.18653/v1/2023.acl-long.256.

[66] Yafu Li, Yongjing Yin, Jing Li, and Yue Zhang. Prompt-driven neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2579–2590, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.203. URL https://aclanthology.org/2022.findings-acl.203.

[67] Yichuan Li, Xiyao Ma, Sixing Lu, Kyumin Lee, Xiaohu Liu, and Chenlei Guo. MEND: meta demonstration distillation for efficient and effective in-context learning. *CoRR*, abs/2403.06914, 2024. doi: 10.48550/ARXIV.2403.06914. URL https://doi.org/10.48550/arXiv.2403.06914.

[68] Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19565–19594. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23l.html.

[69] Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual generative language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9019–9052. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.616. URL https://doi.org/10.18653/v1/2022.emnlp-main.616.

[70] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In Eneko Agirre, Marianna Apidianaki, and Ivan Vulic, editors, *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, 2022. doi: 10.18653/V1/2022.DEELIO-1.10. URL https://doi.org/10.18653/v1/2022.deelio-1.10.

[71] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *CoRR*, abs/2307.03172, 2023. doi: 10.48550/ARXIV.2307.03172. URL https://doi.org/10.48550/arXiv.2307.03172.

[72] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):195:1–195:35, 2023. doi: 10.1145/3560815. URL https://doi.org/10.1145/3560815.

[73] Sheng Liu, Haotian Ye, Lei Xing, and James Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering, 2024.

[74] Xiaodong Liu, Kevin Duh, and Jianfeng Gao. Stochastic answer networks for natural language inference. *arXiv preprint arXiv:1804.07888*, 2018.

[75] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1441. URL https://aclanthology.org/P19-1441.

[76] Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, and Wei Lu. Let's learn step by step: Enhancing in-context learning ability with curriculum learning, 2024.

[77] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR, 2023. URL https://proceedings.mlr.press/v202/longpre23a.html.

[78] Renze Lou, Kai Zhang, Jian Xie, Yuxuan Sun, Janice Ahn, Hanzi Xu, Yu su, and Wenpeng Yin. MUFFIN: Curating multi-faceted instructions for improving instruction following. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1vrS1zwekw.

[79] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *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 8086–8098. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.556. URL https://doi.org/10.18653/v1/2022.acl-long.556.

[80] X Ma. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

[81] Bill MacCartney and Christopher D. Manning. Modeling semantic containment and exclusion in natural language inference. In Donia Scott and Hans Uszkoreit, editors, *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL https://aclanthology.org/C08-1066/.

[82] Arvind Mahankali, Tatsunori B. Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *CoRR*, abs/2307.03576, 2023. doi: 10.48550/ARXIV.2307.03576. URL https://doi.org/10.48550/arXiv.2307.03576.

[83] Costas Mavromatis, Balasubramaniam Srinivasan, Zhengyuan Shen, Jiani Zhang, Huzefa Rangwala, Christos Faloutsos, and George Karypis. Which examples to annotate for in-context learning? towards effective and efficient selection. *CoRR*, abs/2310.20046, 2023. doi: 10.48550/ARXIV.2310.20046. URL https://doi.org/10.48550/arXiv.2310.20046.

[84] Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Noisy channel language model prompting for few-shot text classification. In *Proc. of ACL*, pages 5316–5330, Dublin, Ireland, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.365.

[85] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 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, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.naacl-main.201.

[86] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11048–11064. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.759. URL https://doi.org/10.18653/v1/2022.emnlp-main.759.

[87] Marvin Minsky. Commonsense-based interfaces. *Communications of the ACM*, 43(8):66–73, 2000.

[88] Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. Reframing instructional prompts to GPTk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.50. URL https://aclanthology.org/2022.findings-acl.50.

[89] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.244. URL https://aclanthology.org/2022.acl-long.244.

[90] Shikhar Murty, Pang Wei Koh, and Percy Liang. ExpBERT: Representation engineering with natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2106–2113, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.190. URL https://aclanthology.org/2020.acl-main.190.

[91] Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA, 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1198.

[92] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*, 2019.

[93] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *CoRR*, abs/2209.11895, 2022. URL https://doi.org/10.48550/arXiv.2209.11895.

[94] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

[95] Long Ouyang, Jeffrey 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. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

[96] Boyuan Pan, Yazheng Yang, Zhou Zhao, Yueting Zhuang, Deng Cai, and Xiaofei He. Discourse marker augmented network with reinforcement learning for natural language inference. *arXiv preprint arXiv:1907.09692*, 2019.

[97] Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:258740972.

[98] Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8298–8319. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.527. URL https://doi.org/10.18653/v1/2023.findings-acl.527.

[99] Nita Patil, Ajay Patil, and BV Pawar. Named entity recognition using conditional random fields. *Procedia Computer Science*, 167:1181–1188, 2020.

[100] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *ArXiv preprint*, abs/2304.03277, 2023. URL https://arxiv.org/abs/2304.03277.

[101] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[102] Jonathan Pilault, Amine Elhattami, and Christopher Pal. Conditionally adaptive multi-task learning: Improving transfer learning in nlp using fewer parameters & less data. *arXiv preprint arXiv:2009.09139*, 2020.

[103] Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. GrIPS: Gradient-free, edit-based instruction search for prompting large language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3845–3864, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.277.

[104] Chengwei Qin, Aston Zhang, Anirudh Dagar, and Wenming Ye. In-context learning with iterative demonstration selection. *CoRR*, abs/2310.09881, 2023. doi: 10.48550/ARXIV.2310.09881. URL https://doi.org/10.48550/arXiv.2310.09881.

[105] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 2019.

[106] A Rakhlin. Convolutional neural networks for sentence classification. *GitHub*, 6:25, 2016.

[107] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.naacl-main.191.

[108] Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.92. URL https://aclanthology.org/2021.emnlp-main.92.

[109] Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2439–2455, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.187. URL https://aclanthology.org/2022.findings-naacl.187.

[110] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.20. URL https://aclanthology.org/2021.eacl-main.20.

[111] Timo Schick and Hinrich Schütze. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 390–402, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.32. URL https://aclanthology.org/2021.emnlp-main.32.

[112] Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Wen tau Yih, and Mike Lewis. In-context pretraining: Language modeling beyond document boundaries. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=LXVswInHOo.

[113] Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woomyoung Park, Jung-Woo Ha, and Nako Sung. On the effect of pretraining corpora on in-context learning by a large-scale language model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5168–5186, Seattle, United States, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.naacl-main.380.

[114] Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen, and He He. Measuring inductive biases of in-context learning with underspecified demonstrations. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11289–11310. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.632. URL https://doi.org/10.18653/v1/2023.acl-long.632.

[115] Thoudam Doren Singh, Kishorjit Nongmeikapam, Asif Ekbal, and Sivaji Bandyopadhyay. Named entity recognition for manipuri using support vector machine. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2*, pages 811–818, 2009.

[116] Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. An information-theoretic approach to prompt engineering without ground truth labels. In *Proc. of ACL*, pages 819–862, Dublin, Ireland, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.60.

[117] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.

[118] Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=qY1hlv7gwg.

[119] Cong Sun, Zhihao Yang, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. Biomedical named entity recognition using bert in the machine reading comprehension framework. *Journal of Biomedical Informatics*, 118:103799, 2021.

[120] Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. *ArXiv preprint*, abs/2201.03514, 2022. URL https://arxiv.org/abs/2201.03514.

[121] Zijun Sun, Chun Fan, Qinghong Han, Xiaofei Sun, Yuxian Meng, Fei Wu, and Jiwei Li. Self-explaining structures improve nlp models. *arXiv preprint arXiv:2012.01786*, 2020.

[122] Chuanqi Tan, Furu Wei, Wenhui Wang, Weifeng Lv, and Ming Zhou. Multiway attention networks for modeling sentence pairs. In *IJCAI*, pages 4411–4417, 2018.

[123] Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4645–4657. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.284. URL https://doi.org/10.18653/v1/2023.findings-acl.284.

[124] Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. Compare, compress and propagate: Enhancing neural architectures with alignment factorization for natural language inference. *arXiv preprint arXiv:1801.00102*, 2017.

[125] Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. UL2: unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=6ruVLB727MC.

[126] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. doi: 10.48550/arXiv.2302.13971. URL https://doi.org/10.48550/arXiv.2302.13971.

[127] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. doi: 10.48550/ARXIV.2307.09288. URL https://doi.org/10.48550/arXiv.2307.09288.

[128] Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9840–9855. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.609. URL https://doi.org/10.18653/v1/2023.emnlp-main.609.

[129] Liwen Wang, Rumei Li, Yang Yan, Yuanmeng Yan, Sirui Wang, Wei Wu, and Weiran Xu. InstructionNER: A Multi-Task Instruction-Based Generative Framework for Few-Shot NER. *ArXiv preprint*, abs/2203.03903, 2022. URL https://arxiv.org/abs/2203.03903.

[130] Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*, 2021.

[131] Xinyi Wang, Wanrong Zhu, and William Yang Wang. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *arXiv preprint arXiv:2301.11916*, 2023.

[132] Yaqing Wang and Quanming Yao. Few-shot learning: A survey. *CoRR*, abs/1904.05046, 2019. URL http://arxiv.org/abs/1904.05046.

[133] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Benchmarking Generalization via In-context Instructions on 1,600+ Language Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, 2022.

[134] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5085–5109. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.340. URL https://doi.org/10.18653/v1/2022.emnlp-main.340.

[135] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13484–13508. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.754. URL https://doi.org/10.18653/v1/2023.acl-long.754.

[136] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 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, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

[137] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://openreview.net/forum?id=yzkSU5zdwD.

[138] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[139] Jerry W. Wei, Le Hou, Andrew K. Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc V. Le. Symbol tuning improves in-context learning in language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 968–979. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.61. URL https://doi.org/10.18653/v1/2023.emnlp-main.61.

[140] Jerry W. Wei, Le Hou, Andrew K. Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc V. Le. Symbol tuning improves in-context learning in language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 968–979. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.61. URL https://doi.org/10.18653/v1/2023.emnlp-main.61.

[141] Orion Weller, Nicholas Lourie, Matt Gardner, and Matthew E. Peters. Learning from task descriptions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1361–1375, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.105. URL https://aclanthology.org/2020.emnlp-main.105.

[142] Noam Wies, Yoav Levine, and Amnon Shashua. The learnability of in-context learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/73950f0eb4ac0925dc71ba2406893320-Abstract-Conference.html.

[143] Patrick H Winston. Learning and reasoning by analogy. *Communications of the ACM*, 23 (12):689–703, 1980.

[144] Hui Wu and Xiaodong Shi. Adversarial soft prompt tuning for cross-domain sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2438–2447, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.174. URL https://aclanthology.org/2022.acl-long.174.

[145] Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1423–1436. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.79. URL https://doi.org/10.18653/v1/2023.acl-long.79.

[146] Congying Xia, Wenpeng Yin, Yihao Feng, and Philip Yu. Incremental few-shot text classification with multi-round new classes: Formulation, dataset and system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1360, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.106. URL https://aclanthology.org/2021.naacl-main.106.

[147] Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

[148] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

[149] Benfeng Xu, Quan Wang, Zhendong Mao, Yajuan Lyu, Qiaoqiao She, and Yongdong Zhang. k nn prompting: Learning beyond the context with nearest neighbor inference. In *International Conference on Learning Representations*, 2023.

[150] Haike Xu, Zongyu Lin, Jing Zhou, Yanan Zheng, and Zhilin Yang. A universal discriminator for zero-shot generalization. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki

Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10559–10575. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.589. URL https://doi.org/10.18653/v1/2023.acl-long.589.

[151] Hanzi Xu, Slobodan Vucetic, and Wenpeng Yin. OpenStance: Real-world zero-shot stance detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–324, Abu Dhabi, United Arab Emirates (Hybrid), 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.conll-1.21.

[152] Kai Xu, Zhenguo Yang, Peipei Kang, Qi Wang, and Wenyin Liu. Document-level attention-based bilstm-crf incorporating disease dictionary for disease named entity recognition. *Computers in biology and medicine*, 108:122–132, 2019.

[153] Steve Yadlowsky, Lyric Doshi, and Nilesh Tripuraneni. Pretraining data mixtures enable narrow model selection capabilities in transformer models. *CoRR*, abs/2311.00871, 2023. doi: 10.48550/ARXIV.2311.00871. URL https://doi.org/10.48550/arXiv.2311.00871.

[154] Jinghan Yang, Shuming Ma, and Furu Wei. Auto-icl: In-context learning without human supervision. *CoRR*, abs/2311.09263, 2023. doi: 10.48550/ARXIV.2311.09263. URL https://doi.org/10.48550/arXiv.2311.09263.

[155] Zhe Yang, Damai Dai, Peiyi Wang, and Zhifang Sui. Not all demonstration examples are equally beneficial: Reweighting demonstration examples for in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13209–13221. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.880. URL https://doi.org/10.18653/v1/2023.findings-emnlp.880.

[156] Qinyuan Ye and Xiang Ren. Learning to generate task-specific adapters from task description. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 646–653, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.82. URL https://aclanthology.org/2021.acl-short.82.

[157] Seonghyeon Ye, Hyeonbin Hwang, Sohee Yang, Hyeongu Yun, Yireun Kim, and Minjoon Seo. Investigating the effectiveness of task-agnostic prefix prompt for instruction following. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19386–19394. AAAI Press, 2024. doi: 10.1609/AAAI.V38I17.29909. URL https://doi.org/10.1609/aaai.v38i17.29909.

[158] Da Yin, Xiao Liu, Fan Yin, Ming Zhong, Hritik Bansal, Jiawei Han, and Kai-Wei Chang. Dynosaur: A dynamic growth paradigm for instruction-tuning data curation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4031–4047. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.245. URL https://doi.org/10.18653/v1/2023.emnlp-main.245.

[159] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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)*, pages 3914–3923, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1404. URL https://aclanthology.org/D19-1404.

[160] Wenpeng Yin, Jia Li, and Caiming Xiong. ConTinTin: Continual learning from task instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3062–3072, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.218. URL https://aclanthology.org/2022.acl-long.218.

[161] Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. Ground-truth labels matter: A deeper look into input-label demonstrations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2422–2437. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.155. URL https://doi.org/10.18653/v1/2022.emnlp-main.155.

[162] Yiming Zhang, Shi Feng, and Chenhao Tan. Active example selection for in-context learning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9134–9148. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.622. URL https://doi.org/10.18653/v1/2022.emnlp-main.622.

[163] Yiming Zhang, Shi Feng, and Chenhao Tan. Active example selection for in-context learning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9134–9148. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.622. URL https://doi.org/10.18653/v1/2022.emnlp-main.622.

[164] Zhuosheng Zhang, Yuwei Wu, Zuchao Li, and Hai Zhao. Explicit contextual semantics for text comprehension. *arXiv preprint arXiv:1809.02794*, 2018.

[165] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. Semantics-aware bert for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9628–9635, 2020.

[166] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang, editors, *Proc. of ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR, 2021. URL http://proceedings.mlr.press/v139/zhao21c.html.

[167] Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2856–2878, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.244. URL https://aclanthology.org/2021.findings-emnlp.244.

[168] GuoDong Zhou and Jian Su. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, pages 473–480, 2002.

[169] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=92gvk82DE-.

A 中文唐诗生成实验结果展示

中文唐诗生成是 NLP Beginner 的 task5。这里我们对我们生成的结果做简单展示（表9）。其中，五言只一句话五个字，七言指一句话七个字。绝句指诗中一共四个句子，律诗指诗中一共八个句子。其中，以逗号或句号为结尾作分隔符的是一个句子。

方法	格式要求	唐诗展示
LSTM	随机	半，虚薄薄，执
	五言绝句	雪梅初初暖，更歌歌歌诗。 汝汝飞飞去，犹犹犹忆空。
	七言绝句	空空歌诗诗诗诗，惟称称称称身身。 清夜夜夜夜夜夜，虚虚虚空空空歌。
	五言律诗	汝汝气气气，嵩空空空鞍。 美美人人世，犹忆忆忆空。 歌歌诗诗诗，犹犹犹犹犹。 但但清渭渭，式式歌歌歌。
	七言律诗	君君不见鞴奏再，空空空空美美人。 美人弃犹犹世世，顾我迟迟迟迟迟。 汝汝清清清清清，嵩空空空雨雨雨。 式瞻瞻奏谁谁谁，空空空属属男儿。
	春夏秋冬藏头绝句	春雨雨气薄，夏君不不见。 秋烂烂死死，冬疑疑肠肠。
	上下左右东南西北藏头律诗	上隐隐隐居居间，下分戚里犹唱唱。 左全诗诗诗诗间，右肠肠肠肠箫箫。 东庙庙风吹吹清，南清清夜夜雨雨。 西江江湖空空空，北退退退箫箫箫。
GPT2-Chinese-poem	随机	得蓬蒿径，欣瞻霁月乡。秋声连海树，晓影乱萤光。 有帛民犹困，无田寇亦荒。不堪双鬓白。幸际风云 会，欣荷圣天运，沈冥付子阳。
	五言绝句	华照艳雪凝，步藻池水掐。 以紫菱盘噩，朱铜绣饰佟。
	七言绝句	年随班入金銮懂，归宿春田间开聱。 望诸公十常侍[F]，不及天无路交[F]。
	五言律诗	华旖旎吐芳，蘼含绿意碉。 帘窥翡翠[F]，饰辟邪集喘。 砌蜗涎壁燊，目腐沙雁[F]。 枭怜仆狠倬，恨根毒壅诎。
	七言律诗	自歌兮瑟自呜扯，试觅无声者哉[F]。 欲听风求不得赧，岂真相亦应无[F]。 天风不呼如虎佥，雷鼓空飞石楼[F]。 此意诚谁解者阂，空冥若不能追飞。
	春夏秋冬藏头绝句	春伤行役匆，夏寒日短骜。 秋朝心薄倬，冬如重较螽。
	上下左右东南西北藏头律诗	上昌期肇吉卜悒，下初终纷六壬[F]。 左右衡第一律雯，右朱邸万岁柯[F]。 东西狄徒云何瞪，南君子意若剽[F]。 西未即复南渡淞，北不复东株桃[F]。

Table 9: NLP Beginner task5 中文唐诗生成结果展示