随着大型语言模型（LLM）的快速发展，它们的通用能力已经取得了显著成功。然而，在法律和医学等专业领域的表现仍然受到限制。为了提高模型在这些领域的表现，合成数据以其高质量的输出涌现为一个有前景的解决方案。

现有的方法依赖于大模型的领域先验进行数据合成。然而，领域语料库中嵌入的特定领域知识和语言模式，包括词汇、句法和文体特征，并未被模型先验完全捕捉 (Zhou et al., 2024b)。这一限制促进了利用未标记数据的方法论的发展，未标记数据本身编码了领域特定的特征 (Hamilton et al., 2016; Mudinas et al., 2018)。最近的进展表明，基于未标记数据驱动的数据合成可以提高领域特定数据的质量和任务性能 (Ziegler et al., 2024)，这支持了我们专注于优化未标记数据驱动的方法论。

然而，尽管一些方法专注于使用未标记数据进行领域特定的数据合成，它们仍然存在一些问题。例如，当前领域合成数据生成方法通常依赖于强大的商业模型或大型 LLMs (Taori et al., 2023; Xu et al., 2024; Chen et al., 2024a)。虽然这些模型性能优异，但通常过于昂贵，限制了可访问性 (Bansal et al., 2024)。使用较小的专业模型是另一种选择 (Zelikman et al., 2022; Chen et al., 2024b; Li et al., 2024c)，但它们所涵盖的任务有限，生成对于复杂任务来说过于简单。

为了解决这些限制，我们提出了 AQuilt，这是一种构建数据的框架，它从任何未标记的数据中整合了答案（Answer）、问题（Question）、未标记数据（Unlabeled data）、检查（Inspection）、逻辑（Logic）和任务类型（Task type）。我们训练了一个较小的数据合成模型来合成特定领域的指令调整数据，并降低合成成本。我们引入了逻辑（Logic）和检查（Inspection）来增强模型推理能力并确保合成数据的质量。此外，任务类型（Task type）被扩展以便在训练期间实现对未见任务的泛化。然后，我们使用 DeepSeek-V3 (Liu et al., 2024) 合成了一个高质量的双语数据集（中文和英文），包含 703k 个示例，用于训练低成本、高相关性的数据合成模型。

AQuilt 在涉及两个基础模型和五个任务的实验中，表现出与蒸馏源模型 DeepSeek-V3 相当的性能，同时仅需生产成本的 17 %。此外，与以往的数据合成专用模型相比，例如，虽然 Bonito (Nayak et al., 2024) 性能良好，但其只限于生成需要无标签数据的英语任务，我们的方法在这些相同任务中表现出更优越的性能。进一步的分析证实了结合逻辑和检查的有效性，以及我们的合成数据对下游任务更高的相关性，这进一步有助于模型的高性能表现。

我们的贡献如下：

- 我们提出了 AQuilt，一个可以以低成本从任何未标记数据集中为任何任务合成高相关性数据的框架。通过结合逻辑和检查，我们增强了模型推理能力并提高了数据质量。
- 实验结果表明，AQuilt 在生产成本为 17 % 的情况下，与 DeepSeek-V3 相当。
- 进一步分析表明，逻辑和自我检查有助于提高性能和生成更相关的数据。
- 我们将公开发布我们的数据合成模型、训练数据和代码，以推动更强大的专用 LLM 和数据合成模型的发展。

1 相关工作

无监督数据域合成 最近的研究利用通用大型语言模型的参数化知识进行领域特定的数据合成，从而避免领域无标注数据 (Bao et al., 2023; Deng et al., 2025a; Luo et al., 2025)。领域导向的创新包括 Zhou et al. (2024c) 使用 Self-Instruct (Wang et al., 2023b) 来合成法律问答对，以及 Li et al. (2024b) 使用 GPT-4 生成科学问题。尤其是，Eldan and Li (2023) 通过生成控制词汇的儿童故事展示了受限的领域适应性。现有方法虽然利用强大的大型语言模型直接合成训练数据 (Gilardi et al., 2023; Xie et al., 2024; Hwang et al., 2024)，但主要依赖于商业大型语言模型的现有领域知识 (Achiam et al., 2023; Yang et al., 2023)，其领域数据合成的效率仍然有限 (Palepu et al., 2024)。然而，小模型在没有外部输入的情况下难以利用领域特定知识合成数据 (Deng et al., 2024; Harbola and Purwar, 2025)，这推动了结合小模型与领域特定无标注数据来进行数据合成的研究。

利用未标记数据进行领域数据合成。 因此，我们专注于基于无标签数据的领域数据合成，这更好地平衡了性能和效率。最近，已经提出了专门的方法来整合无标签数据以解决领域差距 (Bartz et al., 2022; Deng et al., 2023, 2025b; Upadhyay et al., 2025)。例如，Nayak et al. (2024) 在包含无标签数据的数据集上训练模型（例如，总结、阅读理解）以进行任务特定的合成。Ziegler et al. (2024) 将检索与上下文学习相结合，以生成需要专业知识的数据。迭代优化技术，如强化自我训练 (Dou et al., 2024) 和回译 (Li et al., 2024c)，使用领域资源进一步改进合成数据。然而，现有解决方案面临两个关键挑

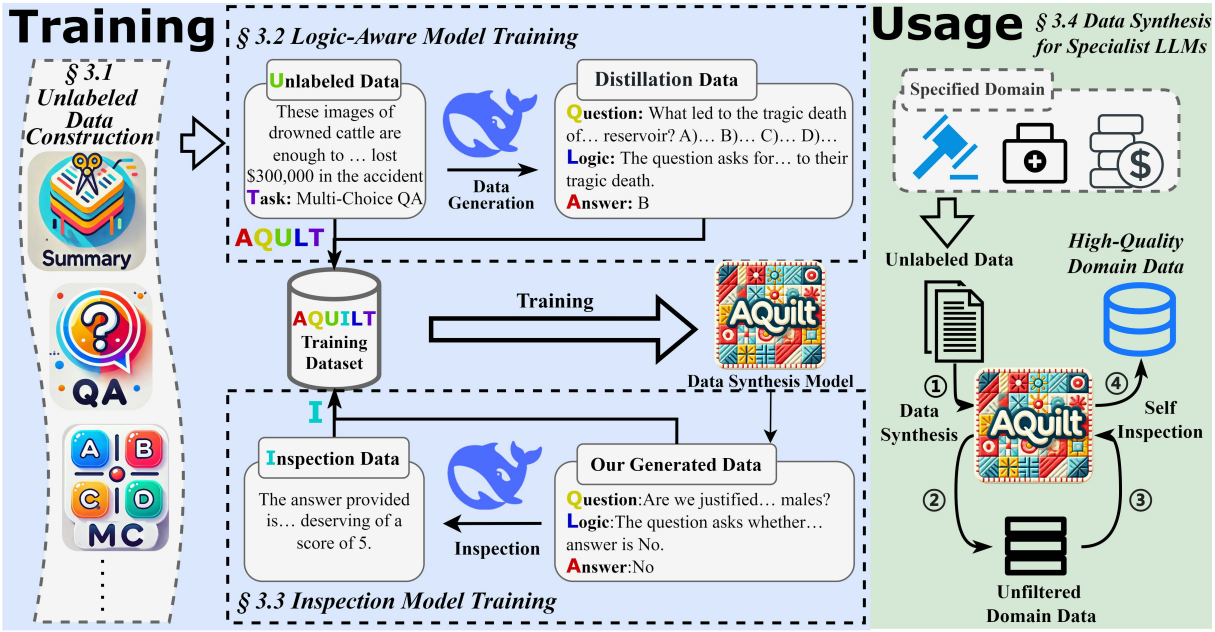

Figure 1: AQuilt 框架提议的概览。左侧展示了我们数据合成模型的训练过程，而右侧展示了经过训练的模型如何自动合成高质量的特定领域数据。合成的数据随后用于训练专用 LLM。

战：（1）依赖商业大型语言模型时成本高且效率低下 (Bansal et al., 2024)，以及（2）专用模型对分布外任务的泛化能力差。例如，像 Chen et al. (2024b) 这样的模型，其在 GPT 生成的种子上训练，缺乏明确定义任务类型的能力。这些限制强调了对能够调和成本效益、领域特异性和任务泛化的框架的紧迫需求。对此，我们提出了一种 7B 参数的数据合成模型，该模型整合了开放域问答以进行任务泛化，同时在无需昂贵的大型语言模型的情况下平衡了效率和质量。

为了开发一种具有跨任务泛化能力的多领域低成本高度相关数据合成模型，我们提出了如图 1 所示的框架。首先，构建一个大型源数据语料库（§**??**）。接下来，我们提取 AQULT 五元组以增强模型的数据合成能力（§1.1）。然后，我们评估生成的数据以获得检查数据，这进一步提高了模型的自我检查能力（§1.2）。最后，利用这些合成数据，我们训练一个稳健的数据合成模型 AQuilt，用于生成高质量的特定领域数据以训练专家级的大型语言模型（§1.3）。

为了使我们的数据合成模型能够合成各种类型的任务，我们遵循 Nayak et al. (2024)，并涵盖了广泛的任务，包括抽取式问答、自然语言推理、多选问答（单答案/多答案）、文本生成、文本摘要、文本分类和自然语言理解。此外，为了增强对下游任务的泛化能力，我们引入了两种额外的任务类型：开放书问答和封闭书问答。由于这些任务的问题是可定制的，因此它们不局限于特定的类别。在为一种新任务类型合成数据时，根据是否需要未标注的数据作为输入，我们将其指定为封闭书问答或开放书问答。此外，我们将新任务的指令作为前缀添加到问题中。这有效地增强了对新领域特定任务类型的数据合成泛化能力。

数据类型。 为了促进数据合成中的多领域泛化，我们聚合了涵盖新闻、百科、评论和多个专业领域的 33 个中英双语数据集的多样化未标记数据，如图 **??** 所示，详情见附录 A。

1.1 逻辑感知模型训练

逻辑感知数据生成 已经证明，结合中间思维推理可以提高 LLM 的性能 (Zelikman et al., 2022)。受此启发，我们将模型的中间推理过程，即逻辑，融入数据合成过程，以促进更有结构的推理过程并提高整体数据质量。具体而言，对于每种任务类型 t，我们随机将其与未标记数据 u 关联，并使用一个强大的商业 LLM，例如 DeepSeek-V3，来生成精炼数据，包括问题 q、逻辑 l 和答案 a。形式上，

$$(a, q, l) = \text{LLM}_{\text{Strong}}^{\text{GenData}}(u, t). \quad (1)$$

此外，我们收集了原始数据集，包括抽取式问答、自然语言推理、多项选择问答（单一答案）和摘要数据集，以增强答案对的多样性，并解决 LLM 在生成抽取式问答数据时面临的挑战。我们进一步提示模型为这些收集的数据集补充缺失的逻辑 l。形式上，

$$l = \texttt{LLM}_{\text{Strong}}^{\text{GenLogic}}(a, q, u, t). \tag{2}$$

收集所有合成数据后，我们获得了数据集 $\mathcal{D}_L = \{(a, q, u, l, t)\}_N$ ，该数据集全面覆盖了 AQuilt 中定义的所有任务类型。

现有的方法，例如 Bonito，通常严重依赖未标记的数据 u 来生成 (q, a) 对，从而引入了为某些不依赖未标记数据的任务合成低相关性数据的风险。与此相反，我们确保我们的合成 (q, a) 对在多选题或闭卷问答任务中保持有意义，而这些任务通常不需要 u 作为输入。为此，我们通过提示工程明确引导模型的偏好。我们还通过识别禁止使用的词语，例如 “the context” 和 “the text”，来过滤掉不符合此标准的情况。这样确保生成的问题在有或没有未标记数据的情况下均适用于下游任务。

此外，为了减少大型语言模型 (Zhou et al., 2024a; Guo et al., 2024) 中的潜在偏差，我们分析词频统计。对于每个任务，我们识别出最常用的词，排除停用词。如果任何词出现在超过 10 % 的数据中，这可能表明有风格偏差。在这种情况下，我们会删除包含这些关键词的问题，以减少其出现频率，并确保最终训练集的多样性和无偏性。

过滤后，我们得到精炼的数据集 $\mathcal{D}_L^{'} = \{(a, q, u, l, t)\}_{N'}$ 。总的数据集大小汇总在表格 1 中。

模型训练。 利用上面获得的大规模数据集 $\mathcal{D}_L^{'}$ ，我们训练数据合成模型。为了使模型能够从未标记数据中合成特定任务的数据，我们使用 u 和 t 作为输入，并训练模型生成 q 、a 和 l 。正式来说：

$$\mathcal{L}_{\text{AQuilt}} = -\sum_{\mathcal{D}_L^{'}}^{j} \log P_\theta(a_j, q_j, l_j \mid u_j, t_j). \tag{3}$$

$[-5pt]$

使用上述损失函数 $\mathcal{L}$ ，我们得到数据合成模型 $\texttt{LLM}_{\text{AQuilt}}$ 。

1.2 检测模型训练

上述模型能够从特定领域的文本中生成任何任务的数据。然而，在某些情况下，生成的数据可能质量较低。为了解决这个问题，我们训练模型以获得自我检查的能力。

为了训练自我检查能力，我们需要收集具有不同质量水平的训练数据。然而，由于由强大的商业 LLMs 合成的数据通常质量较高，我们利用先前训练的 LLM，$\texttt{LLM}_{\text{AQuilt}}$ ，来生成新数据。这确保了合成的数据与我们最终生成过程的分布一致，这对模型训练有利。具体来说，

Task Type	English	Chinese
Extractive QA	16k	16k
Natural Language Inference	49k	33k
Multi-Choice QA (Single Answer)	49k	49k
Multi-Choice QA (Multiple Answers)	28k	31k
Text Generation	33k	33k
Text Summarization	49k	43k
Text Classification	33k	33k
Natural Language Understanding	32k	32k
Open-Book QA	33k	31k
Closed-Book QA	33k	33k
Self-Inspection	7k	7k
Total	362k	341k

Table 1: 来自不同任务的生成训练数据数量。总共收集了 703k 的数据，涵盖了英语和中文。

对于每个任务 t ，我们随机抽样 u 并将其输入到训练的模型 $\texttt{LLM}_{\text{AQuilt}}$ 中，以合成数据。随后，我们使用 DeepSeek-V3 对这些样本进行评分。形式化地，

$$\begin{aligned}(a', q', l') =& \texttt{LLM}_{\text{AQuilt}}(u, t),\\ i =& \texttt{LLM}_{\text{Strong}}^{\text{GenInsp}}(a', q', u, l', t).\end{aligned} \tag{4}$$

$[-10pt]$

因此，我们获得了用于训练 AQuilt 自检能力的数据集 $\mathcal{D}_I^{'} = \{(a', q', u, i, l', t)\}_M$ 。

对于训练，我们通过加入一个 LoRA 适配器 (Hu et al., 2022) 来继续微调之前训练的模型 $\texttt{LLM}_{\text{AQuilt}}$ ，形式上记为 $\texttt{LLM}_{\text{AQuilt}}^{\text{LoRA}}$ 。此修改使得模型能够对其自身的指令调优数据集 (a', q', l') 进行评分，该数据集是基于 (u, t) 生成的。形式上，我们使用以下损失函数优化增强了 LoRA 的模型：

$$\mathcal{L}_{\text{AQuilt}}^{\text{LoRA}} = -\sum_{\mathcal{D}_I^{'}}^{j} \log P_{\theta_{\text{LoRA}}}(i_j \mid a'_j, q'_j, u_j, l'_j, t_j). \tag{5}$$

$[-10pt]$

1.3 专用大型语言模型的数据综合

我们使用构建的 AQuilt 模型从未标记的数据中生成高质量的特定领域指令调整数据。然后，我们可以在生成的数据上训练专家模型，以增强其特定领域任务的性能。

领域数据合成。 在为 LLMs 进行下游任务学习时，我们仅使用指定的任务类型 t 和相关领域的未标记数据 u 来合成高质量的特定领域数据。利用我们的数据合成模型，我们可以高效地生成适合给定领域和任务的训练数据。形式上，

$$(a', q', l') = \mathrm{LLM}_{\mathrm{AQuilt}}(u, t). \quad [-10pt] \tag{6}$$

值得注意的是，如果任务类型 t 在训练集中没有被观察到，我们将 t 指定为闭卷问答或开卷问答，具体取决于它是否需要未标记的数据作为输入。此外，我们将新任务的说明作为前缀添加到问题中。

为了确保生成数据的质量，我们基于自检应用过滤。具体来说，我们首先使用训练好的模型生成一个检查得分：

$$i' = \mathrm{LLM}^{\mathrm{LoRA}}_{\mathrm{AQuilt}}(a', q', u, l', t). \tag{7}$$

随后，我们过滤掉低质量的数据。默认情况下，我们移除检查评分为 2 或更低的数据（在 5 分制中）。如果超过% 的数据得分为 2，这表明任务本质上较为简单，我们只移除得分为 1 的数据。结果是，我们获得高质量的训练数据，促进模型高效适应特定领域。

训练专家大语言模型 我们在由 AQuilt 模型合成的高质量领域特定数据上训练目标模型，以增强其在领域特定任务上的性能。

2 实验

如表 1 所总结，我们构建了一个涵盖 10 种任务的双语数据集（EN / ZH），旨在通过多样化覆盖来提高模型泛化能力。为了防止任务主导并确保分布平衡，我们应用了下采样：逻辑训练数据在每个（任务，语言）对被限制为最多 50k 样本，而自检数据对于每种语言每个分数不超过 2k 样本，分数范围为 1 到 5。最终汇总的数据集包含 703k 样本，完整的提示详见附录 B 。

实验在由 8 个 NVIDIA 4090 24GB GPU 组成的 Qwen2.5-7B-Base 上进行。对于上述两种训练过程，我们使用 AdamW 优化器，学习率为 1e-4，批量大小为 32，LoRA 的 r 和 alpha 均设置为 64，训练 2 个周期。

2.1 AQuilt 的评估设置

在这项工作中，我们在不同的下游任务上进行了实验，涵盖了各种任务类型。对于抽取式问答，我们使用 SquadQA (Rajpurkar et al., 2018)，并遵循在线适应设置 (Hu et al., 2023)，该设置教会大型语言模型未标记数据中包含的领域知识。对于是/否问答，我们选择了 PubMedQA (Jin et al., 2019)，这是一个与医学研究论文相关的英语自然语言推理任务。对于多选问答，我们从 CEVAL (Huang et al., 2023) 中选择了八个科目，涵盖了中国从中学到大学的必修课程。对于翻译和开放式问答任务，我们使用了 LexEval (Li et al., 2024a) 中的法律翻译和法律 EssayQA 任务。这些任务跨越不同领域，验证了我们的数据合成模型的跨领域和跨任务能力。

我们使用准确率评估 PubMedQA 和 CEVAL。遵循 Rajpurkar et al. (2018)，我们使用 SQuAD F1 分数来评估 SquadQA 测试数据集。对于翻译和 EssayQA 任务，与 LexEval (Li et al., 2024a) 一样，我们计算生成输出的 Rouge-L 分数。此外，我们计算翻译和 EssayQA 任务的 BERTScore，列在附录 **??** 中，以确保评估指标的鲁棒性。

领域数据生成。 AQuilt 需要领域特定的无标签数据，我们按以下方式获取：SquadQA 使用测试集数据 (Hu et al., 2023)，PubMedQA 使用其原始训练集，CEVAL 收集教科书，而法律任务（翻译和 EssayQA）使用中文 CAIL (china-ai-law challenge, 2024) 和英语 MAUD/UK-Absp (Wang et al., 2023a; Shukla et al., 2022) 数据集。依据 §1.3 生成任务特定的数据，其中未见过的任务如翻译和 EssayQA 通过问题前缀被设计为闭卷 QA。利用 vLLM (Kwon et al., 2023)（温度 =0.7，top_p=0.95，max_length=1024），我们为每个任务合成了 20k 个训练样本。

基线。 我们根据合成训练数据的来源比较不同的基线。对于 “无” 基线，我们直接提示模型进行评估，而不进行任何训练。对于 “TAPT” 基线，我们按照 Gururangan et al. (2020) 进行任务自适应预训练，仅在未标注数据上训练模型。对于 “Bonito” 和 “DeepSeek-V3（带未标注数据）” 基线，我们使用这些模型基于特定领域文本和任务合成数据，然后在合成数据上微调模型。对于 “DeepSeek-V3（带自指令）” 基线，我们通过应用自指令 (Wang et al., 2023b)（一种使模型能够自主创建训练指令的方法）到 DeepSeek-V3 上生成数据，并从 SELF-GUIDE (Zhao et al., 2024) 调整提示以合成特定领域任务数据。对于 “DeepSeek-V3（带自指令 + 未标注数据）” 基线，在表 **??** 中简称为 “DeepSeek-V3（带 SI + UD）”，我们将未标注数据融入自指令数据合成过程中（具体提示见附录 B ）。

为了验证合成数据的效果，我们基于指导模型进行实验，包括 Qwen2.5-7B-Instruct 和 Llama3-8B-Instruct。我们使用 LoRA 对这些模型进行微调，在不同的监督来源上每个任务进行 3 个 epoch。所有其他设置与第 **??** 节中的一致。需要注意的是，我们使用较低的学习率 1e-7，并基于 Qwen2.5 对 CEVAL 进行单 epoch 训练（由于 Qwen2.5 的强基准容易过拟合）以及 TAPT（为了保持指令遵循能力）。对于 TAPT，我们将任务重新格式化为类似 ‘‘请

输出英文的 [Domain] 文本：[Sentence]'' 这样的调优提示，以保持一致性。

2.2 主要结果

正如表 **??** 所示，AQuilt 平均上优于大多数基线，并且可以与使用 DeepSeek-V3（含 SI + UD）的最佳设置媲美。此外，仅依赖于未标记数据的 TAPT 在各个任务中没有显示出显著的改进，突出了标记数据合成的价值。SquadQA 测试合成数据是否有效促进从未标记数据中学习领域特定知识。由于 DeepSeek-V3（含 Self-Instruct）缺乏未标记数据，因此无法进行此设置，其结果标记为 NA。相比之下，AQuilt 在该任务上显著改善了结果，证实了 LLM 在抽取式 QA 任务上的表现不佳，并证明了使用标记数据中的原始 QA（在 §1.1 中介绍）应用于此类任务是合理的。

低成本生成 我们计算不同数据合成和训练方法的成本，以美元为单位。鉴于 DeepSeek-V3 模型的 671B 尺寸使得本地部署不切实际，我们使用其官方 API 并基于总数据合成支出进行成本计算。对于其他来源，我们使用本地 NVIDIA 4090 24GB GPU，通过 Vast AI [1] 的 GPU 租赁价格计算成本。对于生产成本，我们根据所用的 GPU 小时数计算总成本。

AQuilt 在其成本的 17 % 时性能与 DeepSeek-V3 (带 SI + UD) 相当，并且在其成本的 31 % 时其结果优于 DeepSeek-V3 (带未标记数据)，这表明 AQuilt 在数据合成方面具有显著的效率优势。

由于

跨任务泛化。 Bonito 仅支持依赖于未标记数据的英文任务，无法为三个任务生成数据（结果标记为 NA）。相比之下，AQuilt 通过结合中文数据并定义更灵活的任务类型，达到更好的任务广泛性。具体而言，AQuilt 将任务类型指定为闭卷/开卷问答，并使用任务需求作为问题前缀。实验结果表明，AQuilt 在这些任务上始终优于各种基准，突显了其在不同任务间广泛推广的强大能力。

通过比较 DeepSeek-V3（使用 SI + UD）与 DeepSeek-V3（使用 Self-Instruct），显然，即使在强大的 DeepSeek-V3 模型基础上，在领域数据合成过程中加入未标记数据也能进一步提高合成数据质量。这突出显示了使用领域特定的未标记数据进行数据合成的有效性。

[1] https://vast.ai/

Model	SquadQA	CEVAL	Translation	Avg.
AQuilt	**40.89**	**63.16**	**34.50**	**46.18**
w/o Logic	40.68	59.64	33.61	44.64
w/o Self-Inspection	40.00	60.95	34.22	45.06
w/ Low-Quality	39.81	59.22	33.15	44.06

Table 2: 逻辑和自检的消融结果。我们独立地去除逻辑和自检来观察性能变化。w/ Low-Quality 指的是使用自检来选择低质量数据，与主要实验设置相反。

3 分析

我们提供了全面的分析来展示我们方法的效果及其成功的潜在原因。考虑到实施成本，除非另有说明，实验仅基于从 Llama3-8B-Instruct 获得的结果作为基础模型，重点关注 SquadQA、CEVAL 和翻译。

3.1 逻辑与自我检查分析

为了评估逻辑和自我检查的效果，我们在训练过程中独立移除每个组件。结果如表 2 所示。

逻辑的影响。 为了验证逻辑的效果，我们从整个训练流程中移除逻辑组件（没有逻辑），并使用不包含逻辑的数据重新训练 AQuilt，同时保持所有其他设置和实验配置。在随后的数据合成中，不包含任何逻辑。结果显示，移除逻辑后模型性能显著下降，证明了逻辑在整个数据合成框架中的关键作用。

为了评估自检的影响，我们进行了以下评估：低质量数据过滤的最佳设置（AQuilt）、无过滤设置（w/o Self-Inspection），以及仅使用自检识别的低质量数据（w/ Low-Quality）。为确保可比性，在所有设置下应用一致的数据量。结果证实了自检的显著效果。此外，在 w/ Low-Quality 条件下，我们观察到性能有轻微下降，这表明低质量数据的存在对模型性能产生负面影响。幸运的是，当应用逻辑时，总体性能保持在相对较高的水平。

3.2 生成数据的领域相关性

我们证明了我们的方法通过生成与目标领域高度相关且噪声更低的数据，实现了卓越的性能。为此，我们选择了 CEVAL、Translation 和 SquadQA 作为三个不同的领域任务，每个任务都有 2k 个合成数据样本。使用 Qwen2.5-7B，我们计算了由 DeepSeek-V3（使用未标记数据）和 AQuilt 生成的合成问题的句子嵌入，这些问题都是基于未标记数据创建的。

如图 2 所示，我们应用 t-SNE 进行降维并绘制了一个二维散点图。结果表明，AQuilt 生成的数据更加集中且噪声更少。为了定量确认这一点，我们计算了轮廓系数（Silhouette Score），

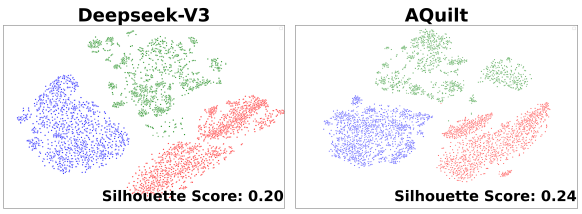

Figure 2: 合成域数据的相关性分析。我们将生成的问题转换为句子向量，并分析在 CEVAL（红色）、Translation（绿色）和 SquadQA（紫色）之间的分布。

Model	SquadQA	CEVAL	Translation	Avg.
DeepSeek-V3	6.90 %	8.18 %	0.00 %	5.23 %
AQuilt	**0.40 %**	**5.15 %**	0.00 %	**1.85 %**

Table 3: 对合成数据与未标记数据的独立性分析。我们评估由 DeepSeek-V3 和 AQuilt 生成的合成问题中高度依赖未标记数据的百分比，揭示它们与下游任务的相关度较低。

该分数反映了数据簇内的一致性和相关性，较高的值表示更强的领域相关性。结果与散点图一致，证明我们的方法生成的数据更加集中，噪声减小，相关性增加。

3.3 相关性感知过滤分析

对于多项选择和闭卷问答任务，合成标注数据必须保持独立于未标注数据；否则，在训练中引入错误关联可能导致与这些任务的低相关性。为了解决这个问题，我们应用了相关性感知的数据过滤（在第 1.1 节中介绍）。

在基于未标记数据的合成生成设置中（表格 **??**），尽管 AQuilt 在各项任务中展现出了一致的优势，但 DeepSeek-V3 (使用未标记数据) 在 CEVAL 上的提升却显得十分有限。为分析其背后的原因，我们检查了生成的问题中依赖未标记数据生成答案的比例，这可能导致幻觉现象。我们从生成的数据集中抽取了 2000 个例子，并使用 GPT-4o 进行评估。如表 3 所示，结果表明即使在提示中明确指出，像 DeepSeek-V3 这样强大的模型依然倾向于生成带有虚假相关性的问题，导致其与后续任务的相关性较低，从而降低整体性能。

3.4 基础模型消融

为了确保改进来源的公平比较，我们在相同设置下（基于相同的未标记数据）进行了对比实验，比较 AQuilt（基于 Qwen2.5-7B 训练）和 Qwen 家族的更大 72B 模型，确认我们的增强是来源于方法论而非基础模型容量。

如表 4 所示，我们的方法在平均性能上优于 72B 模型，同时在 NVIDIA A800 80G 上只需

Source	SquadQA	CEVAL	Translation	Avg.	
				Score	Cost
Qwen2.5-72B	21.19	59.06	**34.82**	38.36	19.92 ×
AQuilt	**40.89**	**63.16**	34.50	**46.18**	**1** ×

Table 4: 与 Qwen2.5-72B-Instruct 的比较。所有缩写均与表 **??** 中的一致。

Model	SquadQA	CEVAL	Translation	Avg.
Bonito	2.43	NA	NA	NA
AQuilt	2.98	4.16	3.58	3.57

Table 5: 使用 GPT-4o 对从由 AQuilt 和 Bonito 合成的训练数据中随机抽取的 1,000 个样本进行评估，以进行这三个测试任务。

要大约 1/20 的 GPU 小时数。整体实验结果与主要实验趋势一致，进一步验证了我们方法的效果，并确认优越的性能并非由于一个更强的基础模型。

为了验证 AQuilt 生成的问题的质量并与同等级模型如 Bonito 进行比较，我们在表 5 中展示了得分结果（没有经过自检过滤）。这些结果是通过使用 GPT-4o（温度 =0.7, top_p=0.95）评估各个任务中 AQuilt 和 Bonito 合成的训练数据中随机抽取的 1,000 个样本得到的。GPT-4o 使用的提示词与附录 B 中展示的完全一致，评分范围从 1 到 5 分。

基于我们给 GPT-4o 提供的提示，得分 2 分即符合基本质量要求。如表 5 所示，大多数任务中，AQuilt 合成的数据的平均得分超过 3 分。然而，Bonito 合成的数据得分较低。同时，由于 Bonito 无法为 CEVAL 和翻译任务合成数据，这些被标记为 NA。这表明大多数由 AQuilt 合成的数据质量相对较高。对于像 SquadQA 和翻译这样相对简单的任务，GPT-4o 倾向于给予稍低的分数，这表明领域特定任务的难度在某种程度上会影响最终的评估结果。

4 结论

在本文中，我们提出了 AQuilt，一个用于生成数据的框架，该框架结合了无标签数据中的检查、问题、未标记数据、答案、逻辑和任务类型。具体而言，AQuilt 通过引入逻辑和检查来提高数据合成质量。任务类型的加入，涵盖了开放书籍 QA 和封闭书籍 QA，使得下游数据合成能够进行跨任务泛化。实验结果表明，AQuilt 在任务泛化和性能方面均优于广泛使用的数据合成模型 Bonito。我们的合成数据甚至可与 DeepSeek-V3 相媲美，同时生产成本低于 17 %。进一步分析表明，虽然使用开源 LLM 通常能获得良好结果，但在格式遵从性和下游任务相关性方面存在不足。这强调了专用数据

合成模型的必要性。我们将公开所有训练细节和模型，以鼓励进一步研究。

5

局限性

在这项工作中，我们仅使用 DeepSeek-V3 作为提炼数据的来源。将数据源扩展到包括人工整理的现有训练数据集和由更强大模型生成的数据，可能会提供多样化的训练数据组合。这种扩展可能进一步增强合成数据风格的多样性，并改善下游模型的性能和鲁棒性。

在这项工作中，我们将数据合成模型的语言能力扩展到两种高资源语言：中文和英语。在未来的工作中，我们有兴趣探索该模型在中低资源语言上的表现，在这些语言中，模型可能表现较差且数据可用性更少。此外，我们将研究该模型在遇到训练时未见到的语言时，是否表现出零样本泛化能力。

随着 DeepSeek-R1 (Guo et al., 2025) 和 Kimi-K1.5 (Du et al., 2025) 的引入，出现了更高级的数据合成框架，这些框架使用迭代数据合成、高质量评价和强化学习来逐步生成更强的数据。我们的自检训练框架有潜力推广到这些框架中。我们有兴趣在未来的工作中探索这一设置。

6

伦理声明 我们的工作遵循 ACL 伦理政策，并公开发布代码以便重现。LLM 可能会表现出种族和性别偏见，因此我们强烈建议用户在特定环境中应用模型之前评估潜在的偏见。此外，由于难以控制 LLM 的输出，用户应对因幻觉产生的问题保持警惕。

7

致谢 本研究部分得到了中国国家自然科学基金（资助编号：62206076）、广东省 & T 计划（资助编号：2024B0101050003）、广东省基础与应用基础研究基金（资助编号：2024A1515011491）以及深圳市科技计划（资助编号：ZDSYS20230626091203008, KJZD20231023094700001, KQTD20240729102154066）的支持。我们感谢匿名审稿人和主审稿人提出的有见地的建议。

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Hritik Bansal, Arian Hosseini, Rishabh Agarwal, Vinh Q. Tran, and Mehran Kazemi. 2024. Smaller, weaker, yet better: Training LLM reasoners via compute-optimal sampling. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *CoRR*, abs/2308.14346.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: investigating adversarial human annotation for reading comprehension. *Trans. Assoc. Comput. Linguistics*, 8:662–678.

Christian Bartz, Hendrik Raetz, Jona Otholt, Christoph Meinel, and Haojin Yang. 2022. Synthesis in style: Semantic segmentation of historical documents using synthetic data. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3878–3884.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2024a. Alpagasus: Training a better alpaca with fewer data. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yongrui Chen, Haiyun Jiang, Xinting Huang, Shuming Shi, and Guilin Qi. 2024b. DoG-instruct: Towards premium instruction-tuning data via text-grounded instruction wrapping. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4125–4135, Mexico City, Mexico. Association for Computational Linguistics.

china-ai-law challenge. 2019. Cail2019 - china ai and law challenge 2019.

china-ai-law challenge. 2024. Cail (china ai and law challenge) official website.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for Chinese machine reading comprehension. 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 5883–5889, Hong Kong, China. Association for Computational Linguistics.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. 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 5925–5932, Hong Kong, China. Association for Computational Linguistics.

DataFountain. 2020. Covid-19 government affairs question answering assistant dataset.

Hexuan Deng, Liang Ding, Xuebo Liu, Meishan Zhang, Dacheng Tao, and Min Zhang. 2023. Improving simultaneous machine translation with monolingual data. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 12728–12736. AAAI Press.

Hexuan Deng, Wenxiang Jiao, Xuebo Liu, Jun Rao, and Min Zhang. 2025a. REA-RL: Reflection-Aware Online Reinforcement Learning for Efficient Large Reasoning Models. *arXiv preprint arXiv:2505.19862*.

Hexuan Deng, Wenxiang Jiao, Xuebo Liu, Min Zhang, and Zhaopeng Tu. 2024. Newterm: Benchmarking real-time new terms for large language models with annual updates. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Hexuan Deng, Wenxiang Jiao, Xuebo Liu, Min Zhang, and Zhaopeng Tu. 2025b. DRPruning: Efficient Large Language Model Pruning through Distributionally Robust Optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Zi-Yi Dou, Cheng-Fu Yang, Xueqing Wu, Kai-Wei Chang, and Nanyun Peng. 2024. Re-ReST: Reflection-reinforced self-training for language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15394–15411, Miami, Florida, USA. Association for Computational Linguistics.

Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, et al. 2025. Kimi k1.5: Scaling Reinforcement Learning with LLMs. *arXiv preprint arXiv:2501.12599*.

Ronen Eldan and Yuanzhi Li. 2023. TinyStories: How Small Can Language Models Be and Still Speak Coherent English? *arXiv preprint arXiv:2305.07759*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Giovanni Grano, Andrea Di Sorbo, Francesco Mercaldo, Corrado A Visaggio, Gerardo Canfora, and Sebastiano Panichella. 2017. Android apps and user feedback: a dataset for software evolution and quality improvement. In *Proceedings of the 2nd ACM SIGSOFT international workshop on app market analytics*, pages 8–11.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. Bias in Large Language Models: Origin, Evaluation, and Mitigation. *arXiv preprint arXiv:2411.10915*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Institute for Frontier Information Haihua and Institute for Interdisciplinary Information Sciences Tsinghua. 2021. 2021 hai hua ai competition.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas. Association for Computational Linguistics.

Chitranshu Harbola and Anupam Purwar. 2025. Knowslm: A framework for evaluation of small language models for knowledge augmentation and humanised conversations. *arXiv preprint arXiv:2504.04569*.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, and et al. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972, Lisbon, Portugal. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,

and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Nathan Hu, Eric Mitchell, Christopher Manning, and Chelsea Finn. 2023. Meta-learning online adaptation of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4418–4432, Singapore. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *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*.

Hyeonbin Hwang, Doyoung Kim, Seungone Kim, Seonghyeon Ye, and Minjoon Seo. 2024. Self-explore: Enhancing mathematical reasoning in language models with fine-grained rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1444–1466, Miami, Florida, USA. Association for Computational Linguistics.

Su Jianlin. 2017. Baidu's chinese question-answering dataset webqa.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. 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 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Pluto Junzeng. 2020. ChineseSquad: Chinese Read Comprehension Dataset.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8082–8090. AAAI Press.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024a. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 25061–25094. Curran Associates, Inc.

Sihang Li, Jin Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. 2024b. ScilitLLM: How to adapt LLMs for scientific literature understanding. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. 2024c. Self-alignment with instruction backtranslation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 58–62, Hong Kong, China. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, Yansong Tang, and Dongmei Zhang. 2025. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. In *The Thirteenth International Conference on Learning Representations*.

Andrius Mudinas, Dell Zhang, and Mark Levene. 2018. Bootstrap domain-specific sentiment classifiers from unlabeled corpora. *Transactions of the Association for Computational Linguistics*, 6:269–285.

Nihal Nayak, Yiyang Nan, Avi Trost, and Stephen Bach. 2024. Learning to generate instruction tuning datasets for zero-shot task adaptation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12585–12611, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

NLPCC. 2017. NLPCC 2017 Shared Task Data.

Anil Palepu, Vikram Dhillon, Polly Niravath, Wei-Hung Weng, Preethi Prasad, Khaled Saab, Ryutaro Tanno, Yong Cheng, Hanh Mai, Ethan Burns, et al. 2024. Exploring large language models for specialist-level oncology care. *arXiv preprint arXiv:2411.03395*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8722–8731. AAAI Press.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Chih-Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. DRCD: a chinese machine reading comprehension dataset. *CoRR*, abs/1806.00920.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal case document summarization: Extractive and abstractive methods and their evaluation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1048–1064, Online only. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge dataset and models for dialogue-based reading comprehension. *Trans. Assoc. Comput. Linguistics*, 7:217–231.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. Investigating prior knowledge for challenging chinese machine reading comprehension. *Trans. Assoc. Comput. Linguistics*, 8:141–155.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. QuaRTz: An open-domain dataset of qualitative relationship questions. 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 5941–5946, Hong Kong, China. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Alibaba Cloud Tianchi. 2020. Tianchi competition: Traditional chinese medicine literature question generation challenge.

Alibaba Cloud Tianchi. 2024. "wanchuang cup" traditional chinese medicine question generation challenge dataset.

Ojasw Upadhyay, Abishek Saravankumar, and Ayman Ismail. 2025. Synlexlm: Scaling legal llms with synthetic data and curriculum learning. *arXiv preprint arXiv:2504.18762*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference*

on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 3261–3275.

Steven Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dmitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2023a. MAUD: An expert-annotated legal NLP dataset for merger agreement understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16369–16382, Singapore. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P Lillicrap, Kenji Kawaguchi, and Michael Shieh. 2024. Monte carlo tree search boosts reasoning via iterative preference learning. In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *arXiv preprint arXiv:2309.17421*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *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*.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. *CoRR*, abs/1810.12885.

Chenyang Zhao, Xueying Jia, Vijay Viswanathan, Graham Neubig, and Tongshuang Wu. 2024. Self-guide: Better task-specific instruction following via self-synthetic finetuning. In *First Conference on Language Modeling*.

Ben Zhou, Hongming Zhang, Sihao Chen, Dian Yu, Hongwei Wang, Baolin Peng, Dan Roth, and Dong Yu. 2024a. Conceptual and Unbiased Reasoning in Language Models. *arXiv preprint arXiv:2404.00205*.

Xiaomao Zhou, Qingmin Jia, and Yujiao Hu. 2024b. Advancing general sensor data synthesis by integrating llms and domain-specific generative models. *IEEE Sensors Letters*, 8(11):1–4.

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024c. Lawgpt: A chinese legal knowledge-enhanced large language model. *Preprint*, arXiv:2406.04614.

Ingo Ziegler, Abdullatif Köksal, Desmond Elliott, and Hinrich Schütze. 2024. CRAFT Your Dataset: Task-Specific Synthetic Dataset Generation Through Corpus Retrieval and Augmentation. *arXiv preprint arXiv:2409.02098*.

A 未标记数据构建细节

我们介绍了收集的数据集的来源，包括未标记数据和标记数据的数据集，其中我们收集了 (u, q, a, t) 个元组以确保更高的质量。

无标签数据的来源。 我们收集了来自以下 33 个数据集的未标记数据：CMRC2018 (Cui et al., 2019) 、ChineseSquad (Junzeng, 2020) 、CHIP2020 (Tianchi, 2020) 、CAIL2019 (china-ai-law challenge, 2019) 、DRCD (Shao et al., 2018) 、Covid19QA (DataFountain, 2020) 、WebQA (Jianlin, 2017) 、CMQG (Tianchi, 2024) 、HaihuaAI (Haihua and Tsinghua, 2021) 、C3 (Sun et al., 2020) 、NLPCC (NLPCC, 2017) 、Dureader (He et al., 2018) 、LCSTS (Hu et al., 2015) 、AdversarialDbidaf, AdversarialDroberta, AdversarialDbert (Bartolo et al., 2020) 、ANLI (Nie et al., 2020) 、APPReviews (Grano et al., 2017) 、CosmaQA (Huang et al., 2019) 、Dream (Sun et al., 2019) 、Duorc (Saha et al., 2018) 、Qasc (Khot et al., 2020) 、Quail (Rogers et al., 2020) 、Quartz (Tafjord et al., 2019) 、Quoref (Dasigi et al., 2019) 、RACE (Lai et al., 2017) 、Ropes (Lin et al., 2019) 、SocialQA (Sap et al., 2019) 、Squad (Rajpurkar et al., 2016) 、SuperGLUE (Wang et al., 2019) 、Record (Zhang et al., 2018) 、WikiHop (Welbl et al., 2018) 。

我们从以下 21 个数据集中收集标注数据：CMRC2018、ChineseSquad、CHIP2020、CAIL2019、HaihuaAI、C3、DRCD、Covid19QA、WebQA、CMQG、Dureader、LCSTS、AdversarialDbidaf、AdversarialDroberta、AdversarialDbert、ParaphraseRC、SelfRC、Quoref、Ropes、Squad、Record。在表格中，我们展示了通过 Rouge-L 和 BERTScore 评估的翻译和 EssayQA 任务的结果。可以看出，AQuilt 合成的数据显著提高了 Instruct 模型在这些任务上的表现，其改进幅度可与 DeepSeek-V3 相媲美。

B 提示

我们在下表中展示了我们论文中使用的所有提示，包括：

- 合成训练数据集的提示：该提示用于在 §1.1 中构建的训练数据集。在使用我们的数据集生成下游领域数据时，遵循相同的提示格式。
- 用于抽取式问答的逻辑生成提示：该提示用于在生成抽取式问答任务的逻辑时。
- 用于合成检测训练数据集的提示：该提示用于生成 DeepSeek-V3 的检测数据。
- 不同任务提示和 AQuilt 的自我检查提示：该提示用于使用 AQuilt 生成下游任务数据，涵盖我们定义的所有任务。
- 基于未标记数据的通用 LLM 下游任务生成提示：此提示用于在使用通用 LLM 生成基于未标记数据的数据时。
- 使用自我指令进行通用 LLM 下游任务生成的提示：该提示用于使用自我指令生成通用 LLM 的数据。
- 用于下游任务评估的提示：提示在下游任务的评估中使用。
- 用于合成数据独立性分析的 GPT-4o 提示：该提示用于在分析未标记合成数据的独立性时。

Prompts for Synthetic Training Dataset (Meta Prompts)

You are a professional Q & A pair generation assistant. Your responsibility is to create complete, clear, accurate, and useful { Task Type } Q & A pairs based on the provided text content. You need to deeply analyze the text to create reasonable questions and provide appropriate, detailed answers, ensuring that each Q & A pair is relevant and useful.
I. The { Task Type } questions you create should meet the following requirements:
Requirement 1: { Question Requirement }
Requirement 2: The { Task Type } questions generated can be answered without <text>, and the question provides comprehensive information, complete context, including the core content or key information of <text>. Specifically: The question must explicitly contain the key information of <text> to ensure that the question itself is self-contained. - Do not rely on external context or assume that the user already understands the content of <text>. Avoid using phrases such as according to the above text, in the context, above content, or based on the information provided;
Requirement 3: The { Task Type } questions generated should be as complex as possible, requiring multi-step reasoning to determine the final answer;
Requirement 4: The generated answer is accurate and error-free, based on credible facts and data from the provided text;
Requirement 5: The generated answer is complete, not only selecting the correct option but also providing explanation and thinking steps to exclude other distractors;
Requirement 6: { Thought Process Requirement }
Requirement 7: When generating Q & A pairs and thought process, assume there is no text as a reference, which means do not include phrases like the text, the context, or the information provided in the Q & A pairs and thought process you create.
II. Please generate a Q & A pair in the following format:
JSON
{
question: { { The question you create } } ,
thought process: { { The thought process you create } } ,
answer: { { The answer you create } }
}

III. Please study the above requirements carefully and create a { Task Type } Q & A pair:

Prompts for Synthetic Training Dataset (Specific Task Content 1/3)

Multi-Choice QA (single answer) :
Task Type: single-choice
Question Requirement: The intent of the single-choice questions generated is clear and the semantics are explicit, including the question and necessary answers as well as distractors;
Thought Process Requirement: The thinking process should include the following steps: 1. Read the question: Understand the provided question. 2. Analyze the options: Assess the relationship and correctness of each candidate with the provided question. 3. Choose the best answer: Select the most accurate and contextually relevant option;

Multi-Choice QA (multi answer) :
Task Type: multi-choice
Question Requirement: The intent of the generated multiple-choice question is clear and the semantics are explicit, including the question and necessary answers (there can be multiple answers that meet the requirements of the question) as well as distractors;
Thought Process Requirement: The thought process should include the following steps: 1. Read the question: Understand the question provided and clarify what is required. 2. Analyze the options: Assess each candidate option's relationship and correctness in relation to the provided question. 3. Choose reasonable answers: Based on the analysis of each option, select all reasonable answers;

Closed-Book QA :
Task Type: closed-book
Question Requirement: The generated questions should be answerable without external knowledge and should provide comprehensive information, complete context, and contain relevant background information; do not generate questions about specific small events, as such questions are meaningless;
Thought Process Requirement: The thought process should include the following steps: 1. Understanding the question: Understand the question and clarify its requirements. 2. Analyzing the question: Analyze relevant information based on your own knowledge without relying on external resources. 3. Formulating a response: Construct a reasonable and accurate answer.

Open-Book QA :
Task Type: open-book
Question Requirement: Open Q & A refers to identifying and extracting specific information segments from the given text to answer the question. The generated question should explicitly include the text needed to answer the question and should not directly use the text provided by the user.
Thought Process Requirement: The generation of the thought chain should include the following steps: 1. Read the text: fully understand the problem and the provided text or paragraph. 2. Identify the relevant parts: locate the specific text segments that contain the answer to the question. 3. Construct the final reply: summarize the answer to the question.

Prompts for Synthetic Training Dataset (Specific Task Content 2/3)

Text Classification :
Task Type: text classification
Question Requirement: The generated classification question should have a clear intent and unambiguous semantics, including the text content and predefined categories.
Thought Process Requirement: The generated thought process should include the following steps: 1. Analyze the content: Examine the content of the <text>, identifying themes, keywords, or other indicative features.2. Map to labels: Match the analyzed features to predefined labels or categories.3. Confirm classification: Verify that the assigned label accurately reflects the content.4. Record the result: Record or output the classification result.

Natural Language Inference :
Task Type: natural language inference
Question Requirement:The question generated can be answered with Yes/No/Maybe.
Thought Process Requirement: The generated thought process should include the following steps:1.Understand the question: Understand the provided question. 2. Analyze the question: Identify which specific parts of the text the question is related to. 3. Logic Reasoning: provide logic reasoning to answer. 4. Provide the best answer: Select yes/no/maybe to answer the question.

Text Generation :
Task Type: text generation
Question Requirement: The text generation problem should have a clear intent and be semantically clear. It must include the user's instructions and the conditions for generation.
Thought Process Requirement: The thought process for generation should include the following steps: 1. Understand the input conditions: Review the user's instructions and conditions to grasp the required scope, tone, and structure. 2. Brainstorm content: Develop key ideas or themes consistent with the input conditions. 3. Generate output: Create a well-structured and coherent response that follows the user's instructions.

Text Summarization :
Task Type: text summarization
Question Requirement: The text summarization problem should have a clear intent and be semantically clear, including the text content and summarization requirements.
Thought Process Requirement: 1. Identify key points: Extract the most important information that represents the overall content of the source material. 2. Organize information: Logically arrange the extracted key points to ensure clarity and coherence. 3. Generate summary: Condense the key points into a concise and clear summary.

Prompts for Synthetic Training Dataset (Specific Task Content 3/3)

Natural Languaga Understanding :
Task Type: natural language understanding
Question Requirement: The generated natural language understanding tasks can specifically include sentiment analysis, intent recognition, entity recognition, part-of-speech tagging, semantic analysis, etc.
Thought Process Requirement: The generated thinking steps should include the following steps: 1. Read the task: Understand the provided task. 2. Analyze the problem: Evaluate the relationship and correctness of the input text in relation to the task. 3. Provide the best answer: Respond with the most accurate and contextually relevant answer.

Logic Generation Prompts for Extractive QA

You are a professional assistant for generating thought process. Your responsibility is to synthesize useful thought process data based on the provided text content and question-answer pairs. You need to deeply analyze the text and construct a logical connection between the question and the answer.
1. The thought process you create should meet the following requirements: Requirement 1: The generated thought process should be rich in content and logically clear, demonstrating how to infer the <Answer> from the <Question>. Requirement 2: The thought process should include the following steps: (1). Read the question: Understand the provided question. (2). Analyze the question: Identify which specific parts of the text the question is related to. (3). Provide the best answer: Select the most relevant content from the text to answer the question.
2. Please generate the thought process in the following format:
JSON
{
“thought_process”: “ { { The thought process you created } } ”
}
3. Please carefully study the above requirements and then create a thought process based on the following <text>, <question>, and <answer> provided by the user:

Prompts for Synthetic Inspection Training Dataset (1/2)

You are an AI instruction and response quality assessment assistant, please score the quality of the user's instruction and response according to the following scoring criteria, and you can refer to the <text> provided by the user to evaluate the correctness of the question and answer pair:
1. The scoring criteria are as follows:
1 point - Low quality, minimal requirements met (Low Level):
- The response is only partially relevant to the question and lacks depth or detail.
- The response contains noticeable grammatical errors, spelling mistakes, or awkward phrasing.
- The response fails to address the user's questions or needs adequately.
- The response provides no additional explanation, context, or background information, leaving the user with limited understanding.
2 points - Basic requirements met (Qualified Level):
- The response is relevant and can basically meet the user's needs.
- The response has correct grammar and no obvious spelling errors.
- The response can solve the user's problem, but the solution may not be comprehensive or in-depth enough.
- The response provides basic explanations and background information, but not in detail.
3 points - Good quality, meeting most requirements (Good Level):
- The response is highly relevant and can well meet the user's needs.
- The response is fluent in grammar, without spelling errors, and clearly expressed.
- The response provides a comprehensive solution that can solve the user's problem and considers possible follow-up issues.
- The response provides detailed explanations and background information, which helps users understand.
4 points - High quality, meeting all requirements and exceeding expectations (Excellent Level):
- The response is highly relevant, not only meeting user needs but also anticipating and resolving potential issues.
- The response has perfect grammar, precise expression, and well-chosen words.
- The response provides an in-depth solution that can comprehensively solve the problem from multiple angles and provides additional useful information or suggestions.
- The response provides in-depth explanations and background information, which helps users gain a deeper understanding of the problem and solution.
5 points - Excellent quality, exceeding all requirements with professional contributions (Outstanding Level):
- The response is highly relevant, not only meeting user needs but also providing solutions beyond expectations.
- The response has impeccable grammar, elegant expression, and precise, impactful wording.
- The response provides an in-depth and professional solution that can solve the problem from a unique perspective and provides extremely valuable additional information or suggestions.

Prompts for Synthetic Inspection Training Dataset (2/2)

```
 - The response provides in-depth explanations and background information,
demonstrating a high level of professionalism and profound understanding of
the issue.
2. Please analysis the quality in the following format:
JSON
{
analysis_steps: { { your analysis for the quality } } ,
score: { { your rate to the qa_pair } }
}

3. Please carefully study the above scoring criteria and strictly follow the
scoring criteria above to score the following <qa_pair> based on the following
<text> provided by the user:
```

Different Task Prompts for AQuilt (1/5)

```
 single choice question answering : """Please generate a single-choice question
from the provided reference materials to help students better grasp the
relevant knowledge: The single-choice question should include a question, four
options labeled A, B, C, and D, one of which is the answer to the question;
At the same time, you also need to generate the thinking steps for solving the
question, as well as the answer to this question.
And output in the following JSON format:
JSON
{ "question": "xxx", "thinking_steps": "xxx", "answer": "xxx" }

""",
multi choice question answering : """Please generate a multiple-choice question
from the references provided to help students better grasp the knowledge:
The multiple-choice question should include a question with multiple options
tags A, B, C, D, E (and so on), one or more of which are the answers to the
questions; At the same time, you also need to generate the thinking steps for
solving the question, as well as the answer to this question.
And output in the following JSON format:
JSON
{ "question": "xxx", "thinking_steps": "xxx", "answer": "xxx" }
"""
```

Different Task Prompts for AQuilt (2/5)

```
close-book question answering : """Please generate a closed-book question and
answer pair from the provided reference materials that do not require reference
text to answer to help students better grasp the relevant knowledge:
This Q & A pair should include a question, and you also need to generate
the thinking steps for solving the question, as well as the answer to this
question.
And output in the following JSON format:
JSON
{ "question": "xxx", "thinking_steps": "xxx", "answer": "xxx" }
""",
open-book question answering : """Please generate an open-book Q & A pair from
the provided reference materials to help students better grasp the relevant
knowledge: This Q & A pair should include a question, and you also need to
generate the thinking steps for solving the question, as well as the answer to
this question. And output in the following JSON format:
JSON
{ "question": "xxx", "thinking_steps": "xxx", "answer": "xxx" }
"""
```

Different Task Prompts for AQuilt (3/5)

```
text summarization : """Please generate a concise summary Q & A pairs of the
provided text to help students better understand the main points:
The summary should capture the key ideas and essential information from the
text. The content you generate should include a question, and you also need to
generate the thinking steps for solving the question, as well as the answer to
this question. And output in the following JSON format:
JSON
{ "question": "xxx", "thinking_steps": "xxx", "answer": "xxx" }
""",
text generation : """Please generate a text-generated Q & A pair based on the
text provided to help students learn:
The resulting text should be well-structured and relevant to the given text.
The content you generate should include a question, and you also need to
generate the thinking steps for solving the question, as well as the answer to
this question. And output in the following JSON format:
JSON
{ "question": "xxx", "thinking_steps": "xxx", "answer": "xxx" }
"""
```

Different Task Prompts for AQuilt (4/5)

```
natural language inference : """Please generate a logical inference question
from the provided reference materials to help students better grasp the
relevant knowledge:
Logical inference questions generally ask whether a judgment or piece of
knowledge is correct, with answers including "yes, no, maybe" three options.
The content you generate should include a question, and you also need to
generate the thinking steps for solving the question, as well as the answer to
this question.
And output in the following JSON format:
JSON
{ "question": "xxx", "thinking_steps": "xxx", "answer": "xxx" }
""",
text classification : """Generate a text classification task based on the text
provided to help students understand the content of the text:
Classifications should be accurate and relevant to the given text.
The content you generate should include a question, and you also need to
generate the thinking steps for solving the question, as well as the answer to
this question.
And output in the following JSON format:
JSON
{ "question": "xxx", "thinking_steps": "xxx", "answer": "xxx" }
"""
```

Different Task Prompts for AQuilt (5/5)

```
extractive question answering : """Please generate an extractive question
answering task based on the provided reference materials to help students
better understand the main points:
The content you generate should include a question, and you also need to
generate the thinking steps for solving the question, as well as the answer to
this question. And output in the following JSON format:
JSON
{ "question": "xxx", "thinking_steps": "xxx", "answer": "xxx" }
""",
natural language understanding : """Please generate a natural language
understanding question (such as sentiment analysis, semantic analysis, entity
recognition, etc.) based on the provided reference materials to help students
better grasp the relevant knowledge:
The content you generate should include a question, and you also need to
provide the thinking steps to solve the question, as well as the answer to the
question. Please output in the following JSON format:
JSON
{ "question":"xxx", "thinking_steps": "xxx", "answer": "xxx" }
"""
```

Self-Inspection Prompts for AQuilt

Please score the quality of the user's instruction and response to help students understand the quality of the question and response based on the provided text. There are 5 levels of quality, which are: 1 point, 2 points, 3 points, 4 points, 5 points. The higher the score, the better the quality.
You'll first need to analyze the quality of the question and response before grading it. And output in the following JSON format:
JSON
{ "analysis_steps": "xxx", "score": "xxx" }

Prompts for General LLM Downstream Task Generation Based on Unlabeled Data (SquadQA)

You are a professional Q & A pair generation assistant. Your responsibility is to create complete, clear, accurate, and useful extractive Q & A pairs based on the provided text.
I. The Q & A pairs you create should meet the following requirements:
Requirement 1: The generated questions should have clear intentions and semantics;
Requirement 2: The generated questions should be answerable without external knowledge; Do not rely on external context or assume that the user already understands the content of <text>.
Requirement 3: The question generated can be answered with the provided reference material.
Requirement 4: The generated answers should be accurate and based on credible facts and data;
Requirement 5: The generated answer can be found in the reference materials.
II. Please generate a Q & A pair in the following format:
JSON
{
"question": " { { The question you create } } ",
"answer": " { { The answer you create(Extracted from reference materials) } } "
}
III. Please study the above requirements carefully and create a extractive Q & A pair based on the <text> provided by the user below:

Prompts for General LLM Downstream Task Generation Based on Unlabeled Data (PubMedQA)

You are a professional Q & A pair generation assistant. Your responsibility is to create complete, clear, accurate, and useful Yes/No Q & A pairs based on the provided text.
I. The Q & A pairs you create should meet the following requirements:
Requirement 1: The generated questions should have clear intentions and semantics;
Requirement 2: The generated questions should be answerable without external knowledge;
Requirement 3: The question generated can be answered with either Yes or No.
Requirement 4: The generated answers should be accurate and based on credible facts and data;
Requirement 5: The generated answer has only two options: Yes/No.
II. Please generate a Q & A pair in the following format:
JSON
{
“question”: “ { { The question you create } } ”,
“answer”: “ { { The answer you create(Yes/No) } } ”
}
III. Please study the above requirements carefully and create a Yes/No Q & A pair based on the <text> provided by the user below:

Prompts for General LLM Downstream Task Generation Based on Unlabeled Data (CEVAL)

You are a professional question-answer pair generation assistant. Your responsibility is to create complete, clear, accurate, and useful single-choice question-answer pairs based on the provided text content. You need to analyze the text in-depth, create reasonable questions, and provide appropriate and detailed answers to ensure that each question-answer pair is relevant and useful.
I. The single-choice questions you create should meet the following requirements:
Requirement 1: The single-choice questions should have clear intentions and be semantically clear, including the question and necessary options as well as distractors;
Requirement 2: The single-choice questions should be answerable without the <text> and the information provided in the question should be comprehensive, with complete context and relevant background information;
Requirement 3: The answers generated should be accurate and based on credible facts and data from the provided text;
Requirement 4: The answers generated should be complete and not omit any necessary information;
II. Please generate the question-answer pairs in the following format:
JSON
{
“question”:“ { { The question you create and the options } } ”,
“answer”:“ { { The correct answer you create } } ”
}
III. After carefully studying the above requirements, please create a single-choice question-answer pair based on the <text> provided below:

Prompts for General LLM Downstream Task Generation Based on Unlabeled Data (Translation)

You are an expert in generating question-and-answer pairs. Your task is to create complete, clear, accurate, and useful closed-book question-and-answer pairs based on the provided text content. You need to analyze the text in depth, formulate reasonable questions, and provide appropriate and detailed answers, ensuring that each question-and-answer pair is relevant and useful.
I. The question-and-answer pairs you create should meet the following requirements:
Requirement 1 The questions should have clear intentions and be semantically clear.
Requirement 2 The questions should be answerable without external knowledge, and the information provided in the question should be comprehensive and contextually complete.
Requirement 3 The questions should be legal translation questions. You need to first determine the language of the given text. If it is in Chinese, the question should be of the Chinese-to-English type. Conversely, if the given text is in English, the question should be of the English-to-Chinese type.
Requirement 4 The type of question you generate can be randomly selected from the following four options: “Please translate the following sentence from the contract into Chinese/English:”, “Please translate the following legal term into Chinese/English:”, “Please translate the following sentence into Chinese/English:”, “Please translate the following legal provision into Chinese/English:”
Requirement 5: The answers you generate should be accurate and based on reliable facts and data.
II. Please generate the question-and-answer pairs in the following format:
JSON
{
“question”:“ { { the question you create } } ”,
“answer”:“ { { the answer you create } } ”
}
III. After carefully studying the above requirements, please create a question-and-answer pair based on the <text> provided by the user below:

Prompts for General LLM Downstream Task Generation Based on Unlabeled Data (EassyQA)

You are a professional question-and-answer pair generation assistant. Your responsibility is to create complete, clear, accurate, and useful question-and-answer pairs based on the provided text content. You need to analyze the text in depth, create reasonable questions, and provide appropriate and detailed answers to ensure that each question-and-answer pair is relevant and useful.
I. The question-and-answer pairs you create should meet the following requirements:
Requirement 1: The generated questions should have clear intentions and be semantically clear.
Requirement 2: The questions should be answerable without external knowledge, and the information provided in the questions should be comprehensive and contextually complete.
Requirement 3: The questions should be legal essay questions, starting with: "Please analyze the following essay question, elaborate on your views, and cite relevant legal provisions and principles. Ensure that you provide sufficient arguments and analysis for each question to clearly demonstrate your deep understanding and flexible application of legal issues."
Requirement 4: The answers generated should be accurate and based on reliable facts and data.
II. Please generate the question-and-answer pairs in the following format:
JSON
{
"question": " { { the question you create } } ",
"answer": " { { the answer you create } } "
}
III. Please carefully study the above requirements and then create a Q & A pair based on the <text> provided by the user.

Prompts for General LLM Downstream Task Generation using Self-Instruct (Input Generation w/o context)

As an InputGenerator , your task is to generate a new [input] based on the [instruction] and some example [input].
Try your best to ensure that the new [input] you generate is distinct from the provided [input] while maintaining a diverse, detailed, precise, comprehensive, and high-quality response. Avoid generating a new [input] that is the same as the provided [input].
Start of instruction
{ { instruction } }
End of instruction
Here are some high-quality [input] for the [instruction]. These [input] can provide you with very strict format requirements .
Below are [N] [input] examples:
{ { Input Examples } }
Please generate 1 [input] based on the examples and requirements:

Prompts for General LLM Downstream Task Generation using Self-Instruct (Input Generation w/ context)

```
As an InputGenerator , your task is to generate a new [input] based on the
[instruction], [context] and some example [input].
Try your best to ensure that the new [input] you generate is distinct from the
provided [input] while maintaining a diverse, detailed, precise, comprehensive,
and high-quality response. Avoid generating a new [input] that is the same as
the provided [input].
Start of instruction
{ { instruction } }
End of instruction
Here are some high-quality [input] and provided [context] for the [instruction].
These [input] can provide you with very strict format requirements .
Below are [N] [input] examples:
{ { Input Examples } }
The provided context content is: { { context } }
Please generate 1 [input] based on the examples, requirements, and the provided
above context:
```

Prompts for General LLM Downstream Task Generation using Self-Instruct (Output Generation)

```
You are an AI question-answering bot, acting as an expert in the field of { {
domain } } . Please refer to the question provided by the user and answer the
question carefully.
User:Please answer the following question:
{ { question } }
```

Prompts For evaluation on Downstream Tasks

SquadQA

Input: { question }

PubMedQA

Input: Context: { context } Based on the context above, please answer the following question: { question }

CEVAL

Input: Below is a multiple-choice question from a Chinese { subject } exam. Please select the correct answer.
{ question }
A. { A } B. { B } C. { C } D. { D }
What is the answer?

Translation

Input: Please complete the following legal translation task and provide the translation directly. Translate the following { text type } into Chinese/English: { legal text } .

EssayQA

Input: Please analyze the following essay question. Elaborate on your views in detail and may cite legal provisions and relevant legal principles. Ensure that you provide sufficient arguments and analysis for each question to clearly demonstrate your profound understanding and flexible application ability of legal issues.
{ material }
Question: { question } .

GPT-4o Prompts for Independence Analysis in Synthetic Data

You are a professional question analysis assistant, responsible for determining whether a question relies on a text for its answer based on the provided question.
Criteria for Judgment:
The question contains some obvious keywords that indicate reliance on a text, such as “the above content,” “according to the text,” “the above text,” “in the text,” “in the passage,” etc.
If the question is about understanding or inquiring about the content of a certain text, then it is also considered a question that relies on a text for its answer.
Formatting Requirements:
Please carefully review the above criteria.
Determine whether the question provided by the user has text dependency.
If it does, please answer directly with ’Yes.’
If it does not, please answer directly with ’No.