TF1-EN-3M：三百万篇用于训练小型开放语言模型的合成道德寓言

Mihai Nad
Babe-Bolyai University
mihai.nadas@ubbcluj.ro

Laura Dioan
Babe-Bolyai University
laura.diosan@ubbcluj.ro

Andreea Tomescu
KlusAI Labs
andreea.tomescu@klusai.com

Andrei Picoran
KlusAI Labs
andrei.piscoran@klusai.com

April 30, 2025

ABSTRACT

寓言故事是传递价值观的一个经典载体，但现代自然语言处理尚缺乏一个将连贯叙事与明确的伦理教训结合起来的大型结构化语料库。我们通过 TF1-EN-3M 填补了这一空白——这是第一个完全由参数不超过 8 B 的指令调优模型生成的英文寓言的公开数据集，包含三百万篇故事。每个故事遵循一个六槽结构（角色 → 特征 → 场景 → 冲突 → 解决方案 → 寓意），通过一个组合提示引擎生成，确保了体裁的忠实性，同时涵盖了广泛的主题空间。一个混合评估流程将 (i) 一个基于 GPT 的评论家与 (ii) 无参考的多样性和可读性度量相结合，评分语法、创造性、寓意明确性和模板遵从性。在十个开放权重候选中，一个 8 B 参数的 Llama-3 变体提供了最佳质量–速度折衷，在单个消费级 GPU（< 24 GB VRAM）上以 ≈ $0.135 生成每 1000 篇寓言。我们在宽松的许可下发布了数据集、生成代码、评估脚本和完整的元数据，允许精确的可重复性和成本基准测试。TF1-EN-3M 为指令遵循、叙事智能、价值观对齐和儿童友好教育 AI 的研究开辟了途径，证明大规模的道德故事叙述不再需要专有的巨型模型。

1 介绍

传授道德教训的故事——寓言——长期以来一直是教授价值观和社会规范的引人入胜的媒介。正如 Guan 等人所指出的，“教授道德是讲故事的最重要目的之一”。传统上，寓言以拟人化的角色为特色，并以明确的道德结尾，将具体事件与抽象的伦理原则相连接。然而，经典的寓言集，如《伊索寓言》，其规模有限，这限制了基于数据的方法来建模道德故事。

近年来，大型语言模型（LLM）的进步为自然语言处理中的合成数据生成开辟了新的可能性。研究人员越来越多地探索 LLM，作为手动标注的经济有效替代方案，生成跨越多个领域的高质量数据集——从分类到对话 [3] 。像 TinyStories [4] 这样的项目已证明，即使是参数少于 1000 万的模型，在经过精心策划的合成数据训练后，也能学会生成连贯的叙述 [5] 。这一进展强调了 LLM 生成语料库在开放模型研究领域，尤其是在低数据状态下的开发潜力。最近的一项全面调查进一步将这些进展置于文本和代码领域中，重点介绍了关键方法，例如基于提示的生成、增强检索管道和强化驱动的精炼 [6] 。

1.1 研究问题

受到以下因素的激励：(i) 缺乏任何大规模、结构化的寓言语料库以进行数据驱动的道德故事叙述，以及 (ii) 需要理解紧凑的开放权重模型是否可以可靠地生成此类叙述，我们调查了：

- 研究问题 1：组合提示扩展方法在使用大语言模型生成多样且高质量的寓言方面有多有效?
- RQ2: 在资源受限条件下，哪种开放权重的大型语言模型最适合生成寓言?

www.xueshuxiangzi.com

受这些问题和新兴趋势的启发，我们开始使用现代开源 LLM 构建一个大规模的寓言数据集。我们的目标是捕捉具有明确教育意图和一致结构的叙述，适用于故事生成、道德推理和指令遵循等应用。关键挑战包括：(a) 通过提示设计系统地涵盖多样化的情境，以及 (b) 在依赖于在消费级硬件上运行的资源有限模型而非昂贵的基于 API 系统的同时，确保生成故事的质量和一致性。

我们的方法采用一个结构化的提示模板，编码经典寓言元素——主角、性格特征、背景、冲突、解决方案和道德。通过组合扩展每个元素的列表，我们生成了数百万个独特的提示，涵盖广泛的道德叙事。

为了确定最适合大规模生成的模型，我们对候选模型（包括 Llama-3、Mistral、DeepSeek 及其他的变体）进行广泛评估。评估沿着两个互补的方向进行：

1. 一个基于大语言模型 (LLM) 的定性评论者，我们使用 GPT-o3-mini 作为一个示例选择，用于评估语法、创造力、道德清晰性和对提示的遵循情况——这反映了最近的研究，该研究表明基于 LLM 的评分与人工评分之间有很强的相关性 [7, 8] 。
2. 一组无参考的多样性和可读性指标，包括自 BLEU [9] 、Distinct-n [10] 和 Flesch 可读性评分 [11] ，它们提供词汇多样性、文本新颖性和可读性方面的独立信号。

这种多视角评估最终使我们选择了 Llama-3.1-8B-Instruct 作为生成 TF1-EN-3M 数据集的模型。虽然其他模型在某些领域取得了略高的分数，但 Llama-3.1-8B 始终在叙述质量和风格简单性之间保持平衡，创作的寓言作品最适合 4 到 7 岁年龄段，因为这一群体自然有限的词汇范围使得这个语料库特别有利于小型下游模型进行参数高效的微调。

1.2 贡献

我们的工作做出了以下关键贡献：

1. TF1-EN-3M 数据集：我们发布了第一个大规模的英文合成寓言数据集（3,000,000 个故事），为故事生成和道德推理的模型训练和评估提供了一种新的资源。
2. 高效高质量生成：我们证明了可以使用相对较小的开放模型（1–8B 参数），在消费者级硬件上大规模生成高质量的、具有指导性的内容。这支持更广泛地访问教学 NLP，并促进具有成本效益的数据集管理 [4] 。
3. 方法创新：我们提出了一种分层评估框架，该框架结合了基于 LLM 的评分和词汇多样性指标，能够进行多准则模型选择，并为寓言生成器奠定基础。

TF1-EN-3M 数据集在 Hugging Face Hub 上以标识符 `klusai/ds-tf1-en-3m` [12] 发布，重新生成和评估该数据集的完整代码在 TinyFabulist 库中可用。

本论文的其余部分组织如下。在第 2 节中，我们形式化了我们的模板驱动提示模式和大规模数据集生成管道。第 3 节介绍了我们的混合大型语言模型评估框架，比较了十个开放权重模型的性能，并将我们的方法置于更广泛的相关工作中。在第 4 节中，我们描述了 TF1-EN-3M 数据集——其结构、元数据架构、生成成本及公开发布情况——在第 5 节中，我们结合原始研究问题反思我们的结果，探索实际应用，并概述对有效性的重要威胁。最后，第 6 节总结了我们的贡献并提出了未来工作的方向。

2 提示设计和数据集生成

我们首先定义一个在叙事理论和先前模板生成工作指导下的结构化提示模式 [13, 14] ，然后构建一个受控的数值空间以从中抽取样本输入。最后，我们指定采样策略和长度限制，以平衡多样性、一致性和计算可行性。

2.1 提示模式和值空间

我们的目标是生成多样化、一致且适合寓言形式的文本。为了在涵盖广泛内容的同时保持结构上的一致性，我们采用了模板驱动的提示设计，这在引导语言模型生成结构化和目的明确的文本方面已显示出有效性 [13, 14] 。具体来说，我们设计的提示模板围绕通过传统叙事结构分析识别出的六个关键寓言元素：

- 角色：主角，通常被描绘成动物或人类原型（例如，狐，樵夫）。
- 特征：一种显著的属性，影响角色行为和故事结果（例如，贪婪、善良、懒惰）。
- 背景：设定叙述在可识别环境中的情景位置（例如，在茂密的森林中，在农场上，在繁忙的小镇中）。
- 冲突：叙述紧张的核心挑战或困境（例如，被某人的诡计骗走了食物，必须在真相与谎言之间做出选择）。

www.xueshuxiangzi.com

- 解决：解决冲突的方法，通常展示道德或伦理判断（例如，角色学习分享，诡计者被揭露）。
- 寓意：明确的课程强调叙述意图（例如，“不要喊狼来除非你是认真的”，“诚实是最好的策略”）。

提示构建： 我们开发了一个提示模板，将这些元素整合到一个结构化的叙事提示中。一般形式如下：

> 创建一个寓言，基于以下元素。自然地将它们编织成一个故事：
> – 主要角色：一个 [特征] [角色]
> – 背景：[情景]
> – 挑战：[冲突]
> – 结果：[解决]
> – 教训：[寓意]

这种明确的结构化确保生成的故事遵循传统的叙事模式，提供了清晰的开头、中间和结尾，以及明确的道德教训，从而直接解决了使生成的故事与指定结果对齐的记录挑战 [15, 16] 。

除了信息填充之外，我们还为提示增加了风格约束，以提升叙事质量并符合寓言的体裁。模板包含了明确的指示，说明模型应该如何实现故事中的每个元素：以生动的场景设置开始，避免使用角色名字（而是提及他们的角色和特质），使用有意义的对话，隐含地展示角色的成长（而不是明说），并以清晰且富有道德启示的结论结束。这些风格提示作为柔性控制，引导语言模型生成连贯且有教育意义的输出，同时保持创作的自由。这种方法借鉴了叙事写作原则，例如 “展示，而不是直白叙述”，并模仿传统寓言的语气和形式，传统寓言往往依靠简明、生动的叙事来传达永恒的道德教训。类似的策略已被证明在神经故事生成中能够改善叙事连贯性和体裁一致性 [15, 16] ，并且与最近基于指令的开放式生成任务控制性研究 [17, 14] 一致。

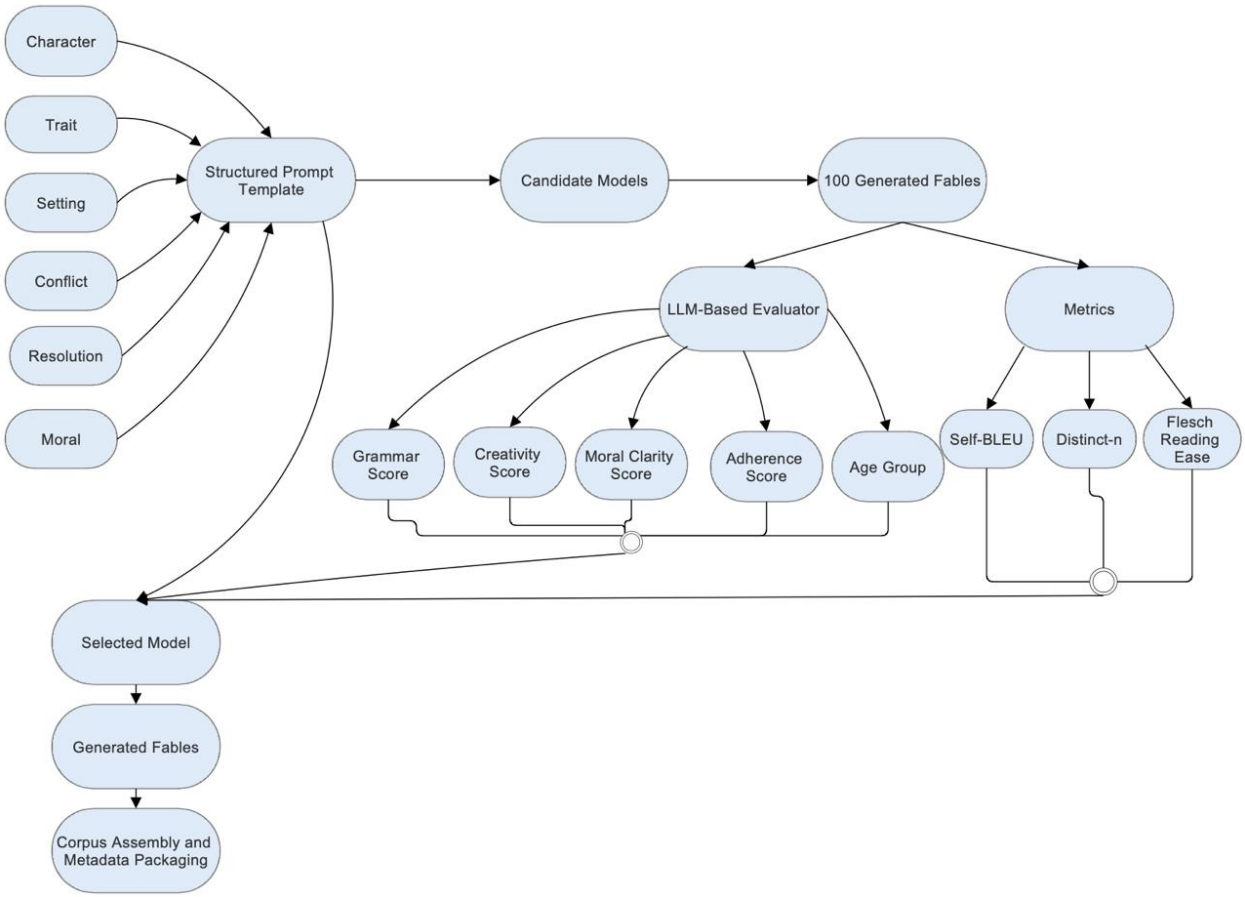

Figure 1: 生成 TF1-EN-3M 的完整流程。

系统消息设计： 除了结构化的叙述模板外，我们还为每个语言模型提供了一个专门的系统消息，将任务框定为道德寓言写作。这一指令设定了对输出质量的期望、特定于体裁的限制以及适合年龄的叙述风格。

www.xueshuxiangzi.com

系统消息强调了三个关键原则：

- 富有想象力且连贯的叙事，
- 读者意识，包括对年轻读者的适宜性，
- 遵循经典寓言格式，由六个核心要素组成：角色、特点、背景、冲突、解决和寓意。

它还定义了五个不同的年龄组（A-E），在评估过程中用于评估目标受众的匹配度。这些受众类别帮助我们基于 LLM 的评论员将每个故事分配到适合的人口群体。

系统消息通过设定语气、伦理明晰和教学目的来补充结构化提示——这是道德故事中的关键元素。它充当全局指令层，确保在百万个独立采样和生成的寓言中保持体裁忠实。

值空间和提示设计变量 我们为提示模板中的每个槽位精心整理了结构化的值集，形成了一个受控的合成生成输入空间。这些集合来源于经典寓言、常见道德主题和创造性扩展的结合，确保了文化相关性和叙述多样性，并由我们的研究团队手动填充。

我们通过引入六个形式参数来抽象输入空间：

$$n,\ m,\ k,\ c,\ r,\ l$$

，其中

- 字符选项的 #
- m = 特征选项的 #,
- k = 设置选项的 #,
- c = 冲突选项的 #
- r = 分辨率选项的 #,
- l = 道德选项的 #。

，因此，总的组合值空间为

$$T \ = \ n \times m \times k \times c \times r \times l.$$

。通过这种方式参数化，我们将模式定义与具体实例化清晰地分开。具体的参数值在第 3 节中指定。

下面我们展示驱动生成器的确切系统消息和提示模板。

系统消息

```
You are a world-class creative assistant that generates captivating and morally-driven fables based
on structured inputs.
Each fable must be:
- Imaginative and coherent.
- Appropriate for a wide audience, including young readers.
- Structured around a classic fable format (character, setting, conflict, resolution, and moral).

Age groups are defined as:
- A: 3 years or under
- B: 4-7 years
- C: 8-11 years
- D: 12-15 years
- E: 16 years or above
```

提示模板

```
Create a fable based on the following elements. Weave them naturally into a story:
- Main Character: a {{trait}} {{character}}
- Setting: a {{setting}} where our story unfolds
- Challenge: {{conflict}}
- Outcome: {{resolution}}
- Teaching: {{moral}}
The fable should:
```


www.xueshuxiangzi.com

```
- Be appropriate for age group B (4-7 years)
- Use simple vocabulary that 4-7 year olds can understand
- Use concrete rather than abstract language
- Begin with vivid scene-setting
- Not use names for the characters, instead use the trait and character
- Include meaningful but simple dialogue
- Show (don't tell) the character's growth
- End with a clear connection to the moral
Keep the story concise but engaging, around 250 words.
```

2.2 评价方法

评估开放式叙事生成是众所周知的具有挑战性的任务。最近的研究探索了使用大型语言模型（LLMs）作为学习评估者，通过提示它们以评分标准驱动的指令来生成多维度评分，以更好地与人类偏好对齐 [18, 7] 。

基于 LLM 的批评者 在 G-Eval 框架 [7] 及类似的总结和对话评估工作 [18] 基础上，我们提示一个指令调整过的 LLM 充当文学评论家。模型接收明确的评分标准（以及在有用时的思维链指导），并按照四个维度为每个寓言打分，分数从 1 到 10 不等：

- 语法 & 风格：语言正确性和句法流畅性，
- 创造力：叙事原创性和创造性，
- 道德清晰性：伦理教训的明确性和相关性，
- 提示依从性：对模板结构和风格限制的忠诚度。

这种方法利用了 LLM 的深刻上下文理解能力，在规模上近似于人的判断，同时效率高于将顶级评估者应用于数百万个例子。

年龄组分类 为确保我们的寓言适合不同的发展阶段，我们还提示同一个大型语言模型将每个故事根据词汇简单性、主题成熟度和句法复杂性分类到五个年龄段之一——A（0-3 岁）、B（4-7 岁）、C（8-11 岁）、D（12-15 岁）、E（16 岁以上）。这反映了微型小说可读性研究中的方法 [4]，帮助我们验证生成内容是否符合预期的教育用途。

无参考指标 除了这些主观评估之外，我们还计算了三个语料库级别的无参考统计数据：

- 自我 BLEU [9]：通过将每个故事与其余故事进行比较来量化集合内的冗余（较低的值表明更大的多样性），
- Distinct-n [10]：衡量独特的 n -grams 的比例（我们报告 $n = 1$ ）作为词汇丰富度的一个代理，
- Flesch 阅读易度 [11]：通过句子和音节计数来评估整体可读性，对适龄理解至关重要。

通过结合基于 LLM 的批评、年龄等级分类和自动化多样性/可读性指标，我们的混合评估流程提供了对寓言质量的强大、多方面的评估——解决了纯参考方法的已知限制，并反映了生成评估的最佳实践 [8] 。

3 大型语言模型的评估与相关工作的比较

3.1 实验设置

为了研究我们的合成寓言在仔细审查下表现如何，我们首先从完整的组合空间

$$T = n \times m \times k \times c \times r \times l,$$

生成了最多 $\texttt{MAX} = 3{,}000{,}000$ 个独特提示，其中每个六个参数 n, m, k, c, r, l 都设置为 100。我们随机均匀抽样，并执行三个温和的保护措施以保持多样性和平衡：

- 唯一性——我们剔除了任何重复的提示；
- 频率筛选——我们对过于常见的冲突-道德配对进行了降采样；
- 覆盖平衡——我们确保每个插槽出现的次数大致相同。

www.xueshuxiangzi.com

每个寓言的最大长度为 1000 个符号，这个长度与传统寓言结构 [19] 很好地对齐。所有使用的模型都在一致的设置下解码——温度设置为 0.7，默认 top-p（即 1.0），基于这些参数禁用了贪婪解码——以确保公平的横向比较 [20]。

我们评估了十个指令微调的、可在消费者 GPU（<24 GB VRAM）上部署的开放权重 LLM：

1. SmolLM2-1.7B-Instruct [21]
2. Aya-23-8B [22]
3. Llama-3.2-1B-指导 [23]
4. Llama-3.1-8B-指令 [24]
5. Llama-3.1-Tulu-3-8B [24]
6. Mistral-7B-Instruct-v0.3 [25]
7. Qwen2.5-7B-Instruct [26]
8. deepseek-llm-7b-chat [27]
9. Phi-3-mini-4k-instruct [28]
10. Falcon3-7B-Instruct [29]

每个模型在相同的解码超参数下为每个提示生成 100 个寓言。

3.2 结果与解释

基于 LLM 的评估 表 1 显示了我们模型在语法、创造力、道德清晰度和提示依从性方面的对比，以及总的平均值、词元数量和推理时间。

Table 1: 基于 LLM 的寓言生成评估（1-10 分制），以及生成元数据。前五个指标列（语法-均值）中的最高值以粗体显示；最低延迟以粗体显示。

Model	Grammar	Creativity	Moral Clarity	Adherence	Mean	Input Tokens	Output Tokens	Latency (s)
Aya-23-8B	7.78	5.75	7.24	5.12	6.47	171.8	500.6	257.89
SmolLM2-1.7B-Instruct	7.79	5.40	6.98	4.81	6.25	174.3	414.7	17.58
Qwen2.5-7B-Instruct	8.28	6.21	8.02	6.81	7.33	182.6	404.2	17.72
Llama-3.1-Tulu-3-8B	8.32	6.97	8.50	7.69	7.87	181.5	368.5	16.91
deepseek-llm-7b-chat	8.04	6.08	7.88	5.72	6.93	189.3	439.8	82.36
Llama-3.1-8B-Instruct	8.42	6.59	8.21	8.18	7.85	181.5	337.6	28.87
Llama-3.2-1B-Instruct	7.87	5.41	6.56	4.98	6.21	181.5	358.5	16.69
Phi-3-mini-4k-instruct	8.10	6.28	7.87	6.61	7.21	0.0	0.0	40.76
Mistral-7B-Instruct-v0.3	8.12	6.31	8.05	6.58	7.26	201.1	426.4	40.70
Falcon3-7B-Instruct	8.29	6.56	8.27	7.06	7.54	186.2	400.4	20.72

结果表明，Llama-3.1-Tulu-3-8B 以总平均最高的 **7.87** 位居榜首，这得益于其在创造力（**6.97**）和道德清晰度（**8.50**）方面的领先得分。紧随其后的是 Llama-3.1-8B-Instruct，在语法（**8.42**）和遵循度（**8.18**）方面表现最强，彰显其与我们的结构化提示良好对齐。Falcon3-7B-Instruct 表现均衡（平均 7.54），拥有高道德清晰度（8.27）和适中的延迟。其他中型模型，如 Mistral-7B-Instruct-v0.3 和 Phi-3-mini-4k-instruct，也取得了可观的平均得分（分别为 7.26 和 7.21）。在效率方面，Llama-3.2-1B-Instruct 实现了最快的推理时间（**16.69** 秒），同时保持合理的质量，展示了其在对延迟敏感或资源受限部署中的适用性。

无参考指标 表 2 引入了三个语料库级别的度量：用于内部多样性的自 BLEU、用于词汇丰富性的 Distinct-1，以及用于可读性的 Flesch 阅读难度。

基于这些指标，我们观察到 Llama-3.1-8B-Instruct 在 Self-BLEU、Distinct-1 和 Flesch Reading Ease 之间展现了一个令人满意的平衡，使其结果既多样化又易于理解。

最后，表 3 显示了每个模型的寓言被认为适合每个年龄段（A–E）的频率。

我们注意到，Llama-3.1-8B-Instruct 及其 Tulu 变体生成了最高比例的 B（4-7 岁）故事，这与我们针对道德和语言简单性 [4] 的目标受众一致。

为了将基于 LLM 的指标和语料库级别的指标整合到一个单一的排名中，我们为每个模型 m 定义一个复合评分 W_m，覆盖七个维度：语法、创造力、道德清晰性、依从性、自 BLEU、Distinct-1 和 Flesch 阅读容易度。我

www.xueshuxiangzi.com

Table 2: 所有被评估模型的非 LLM 文本质量指标（较低的 Self-BLEU 表明更大的多样性；较高的 Distinct-1 表明更丰富的词汇；较高的 Flesch 阅读容易度表明更好的可读性）。最低的 Self-BLEU 和最高的 Distinct-1 及 Flesch 分数被加粗。

Model	SelfBLEU	Distinct1	Flesch Reading Ease
Aya-23-8B	0.361	0.608	73.868
SmolLM2-1.7B-Instruct	0.364	0.567	72.808
Qwen2.5-7B-Instruct	0.390	0.602	80.846
LLaMA-3.1-Tulu-3-8B	0.333	0.659	74.205
deepseek-llm-7b-chat	0.355	0.586	70.731
LLaMA-3.1-8B-Instruct	0.351	0.604	80.071
LLaMA-3.2-1B-Instruct	0.398	0.635	80.832
Phi-3-mini-4k-instruct	0.318	0.651	77.912
Mistral-7B-Instruct-v0.3	0.360	0.634	73.974
Falcon3-7B-Instruct	0.369	0.661	74.379

Table 3: 根据 LLM 分类估算，每个模型生成的寓言在年龄组中的分布（百分比）。年龄组 B 的最高百分比以加粗显示。

Model	Age A	Age B	Age C	Age D	Age E
CohereForAI/aya-23-8B	0.0 %	34.0 %	66.0 %	0.0 %	0.0 %
HuggingFaceTB/SmolLM2-1.7B-Instruct	0.0 %	47.0 %	53.0 %	0.0 %	0.0 %
Qwen/Qwen2.5-7B-Instruct	0.0 %	90.0 %	10.0 %	0.0 %	0.0 %
allenai/Llama-3.1-Tulu-3-8B	0.0 %	71.0 %	29.0 %	0.0 %	0.0 %
deepseek-llm-7b-chat	0.0 %	39.0 %	61.0 %	0.0 %	0.0 %
meta-llama/Llama-3.1-8B-Instruct	0.0 %	92.0 %	8.0 %	0.0 %	0.0 %
meta-llama/Llama-3.2-1B-Instruct	0.0 %	67.0 %	33.0 %	0.0 %	0.0 %
microsoft/Phi-3-mini-4k-instruct	0.0 %	62.0 %	38.0 %	0.0 %	0.0 %
mistralai/Mistral-7B-Instruct-v0.3	0.0 %	47.0 %	53.0 %	0.0 %	0.0 %
tiiuae/Falcon3-7B-Instruct	0.0 %	76.0 %	24.0 %	0.0 %	0.0 %

们将最大的权重赋予依从性，然后是道德清晰性，并在其他五个指标之间平均分配剩余的权重。具体地，让

$$w_{\text{Adh}} = 0.35, \quad w_{\text{Gra}} = w_{\text{Mor}} = 0.20, \quad w_{\text{Cre}} = 0.10, \quad w_{\text{SB}} = w_{\text{D1}} = w_{\text{FRE}} = \frac{0.15}{3} = 0.05.$$

我们首先将每个原始分数 $S_{m,k}$ 标准化为 $\tilde{S}_{m,k} \in [0,1]$ ，通过

$$\tilde{S}_{m,k} = \frac{S_{m,k} - \min_{m'} S_{m',k}}{\max_{m'} S_{m',k} - \min_{m'} S_{m',k}},$$

反向标准化自 BLEU 以使得越低越好。然后

$$\begin{aligned} W_m = {} & w_{\text{Gra}}\, \tilde{S}_{m,\text{Grammar}} + w_{\text{Cre}}\, \tilde{S}_{m,\text{Creativity}} + w_{\text{Mor}}\, \tilde{S}_{m,\text{MoralClarity}} \\ & + w_{\text{Adh}}\, \tilde{S}_{m,\text{Adherence}} + w_{\text{SB}}\, \tilde{S}_{m,\text{Self–BLEU}} \\ & + w_{\text{D1}}\, \tilde{S}_{m,\text{Distinct–1}} + w_{\text{FRE}}\, \tilde{S}_{m,\text{FleschEase}}. \end{aligned}$$

通过构建 $0 \le W_m \le 1$ 。表 4 报告了所有十个模型的加权得分。如图所示，LLaMA-3.1-8B-Instruct 实现了最高的复合得分 **0.891** ，证实其为 TF1-EN-3M 生成的最佳选择。

X_0 我们的方法处于合成数据创建、叙述生成和评估方法等几个活跃研究领域的交集。

首先，使用大型语言模型来生成大规模的合成语料库，作为解决数据稀缺性的一种手段，已经获得了动力。早期的研究表明，对于主观任务的效果是混合的 [30] ，但最近的“由大型语言模型驱动”的流程显示出了显著的规模提升：例如，Persona Hub 使用数十亿的角色模板以空前的规模生成指令风格的数据 [31] ，而自我指令方法则利用 GPT-4 为较小模型引导微调语料库 [32, 5] 。我们的工作遵循这一范式，但特别专注于道德寓言，使用组合式插槽填充而不是角色角色或问答提示来推动内容的多样性。

其次，在故事和寓言生成领域，传统系统依赖启发式规划者或固定的情节模板来强制叙事结构 [33] 。最近，像 TinyStories 这样的项目展示了紧凑的 LLMs（GPT3.5/4）可以使用受控词汇生成数千万个儿童故事，甚至可以通过基于 LLM 的“教师”模型来评估这些故事 [4] 。我们通过针对道德寓言体裁来扩展这项工作——每个故事都围绕一个明确的伦理教训，并提供更加丰富的元数据方案来大规模跟踪生成条件和成本。

www.xueshuxiangzi.com

Table 4: 每个模型的复合加权评分 W_m ，使用 $w_{\text{Adh}} = 0.35$ 、$w_{\text{Gra}} = w_{\text{Mor}} = 0.20$ 、$w_{\text{Cre}} = 0.10$ 、$w_{\text{SB}} = w_{\text{D1}} = w_{\text{FRE}} = 0.05$ 计算。

Model	Grammar	Creativity	Moral Clarity	Adherence	Self-BLEU	Distinct-1	FRE	W
LLaMA-3.1-8B-Instruct	8.42	6.59	8.21	8.18	0.351	0.604	80.071	0.891
LLaMA-3.1-Tulu-3-8B	8.32	6.97	8.50	7.69	0.333	0.660	74.205	0.874
Falcon3-7B-Instruct	8.29	6.56	8.27	7.06	0.369	0.661	74.379	0.729
Qwen2.5-7B-Instruct	8.28	6.21	8.02	6.81	0.390	0.602	80.846	0.640
Phi-3-mini-4k-instruct	8.10	6.28	7.87	6.61	0.318	0.651	77.912	0.608
Mistral-7B-Instruct-v0.3	8.12	6.31	8.05	6.58	0.360	0.634	73.974	0.576
deepseek-llm-7b-chat	8.04	6.08	7.88	5.72	0.355	0.586	70.731	0.392
Aya-23-8B	7.78	5.75	7.24	5.12	0.361	0.608	73.868	0.185
LLaMA-3.2-1B-Instruct	7.87	5.41	6.56	4.98	0.398	0.635	80.832	0.132
SmolLM2-1.7B-Instruct	7.79	5.40	6.98	4.81	0.364	0.567	72.808	0.078

第三，像 STORAL 这样的双语语料库收集了人类创作的道德故事，以研究机器对隐含教训的理解能力 [1] 。STORAL 强调，现成的模型难以将情节事件与道德相匹配，这促使需要专门的训练数据。TF1-EN-3M 通过提供一个合成的大规模平行道德寓言语料库来补充 STORAL，这些寓言既适合微调叙述生成器，也适合在大规模评估中考察道德推理能力。

最终，使用传统指标（BLEU、ROUGE、METEOR）评估开放性叙述时，往往无法体现创意性和主题连贯性，并对合法的词汇变化进行惩罚 [34] 。像 G-Eval 这样的框架已经表明，当给 GPT-4 设定明确的评分标准和思维链提示时，它可以在多个维度上产生与人类一致的评分 [7, 18] 。我们的混合评估管道——结合基于 LLM 的评论员与无参考指标（Self-BLEU、Distinct-n、Flesch Ease）以及年龄组分类——基于这些见解进行构建，提供了一种鲁棒的、多维度的评估，专为寓言创作的教育和结构需求量身定制 [8] 。

通过将合成数据扩展、特定类型叙事生成和混合评估这几个方面结合在一起，我们提出了一种统一的方法，可以以前所未有的规模创造和严格评估道德寓言。

4 TF1-EN-3M 数据集描述和可用性

如上所述，TF1-EN-3M 数据集包括 3,000,000 篇英语寓言故事，每篇都通过结构化提示系统生成并注释了相关的元数据。不同于许多仅提供原始系统输入或最终输出的文本语料库，TF1-EN-3M 以 JSON 行格式保存详细记录，使得提示的具体内容和生成背景的透明检视成为可能。通过结合元数据和叙述，TF1-EN-3M 遵循了类似于其他强调可重复性和道德一致性 [1, 30] 的语料库设计理念。

每个记录包含分为两个主要类别的字段：

4.1

(1) 寓言内容

- 语言：寓言的语言（当前为 `en` ）。
- 提示：描述主题元素（角色、环境、冲突、解决、道德）和风格约束（例如，词数限制，避免使用角色名称）的字符串。这个提示作为输入提供给模型。
- 寓言：由模型生成的完整故事，通常为 1 到 3 段，并以明确的寓意结尾（例如，“诚实是上策。”）。
- 哈希：为每个提示生成一个 SHA-256（安全哈希算法 256 位）加密哈希标识符，用作确保完整性并启用回退机制的手段。

4.2

（2）生成元数据

- llm_name：用于生成的模型的标识符（例如，`tiiuae/Falcon3-7B-Instruct` ）。
- llm_input_tokens, llm_output_tokens：提示和生成寓言的标记计数，对于分析模型效率和冗长性很有用 [35] 。
- llm_inference_time：模型生成输出所需的耗时（以秒为单位），用于进行延迟和吞吐量分析。

www.xueshuxiangzi.com

- host_provider, host_dc_provider, host_dc_location：关于推理基础设施的信息，包括服务提供商和数据中心位置（例如，`eu-west-1` 中的 Hugging Face 推理端点）。
- host_gpu, host_gpu_vram, host_cost_per_hour: 关于 GPU 硬件（类型和显存）以及生成的每小时成本的详细信息，以便于重现性和成本基准测试。
- generation_datetime: 指示故事生成时间的时间戳。
- pipeline_version: 数据生成流程的内部版本，用于在数据集更新时保证可追溯性和可重复性。

发电成本： 生成所有 3,000,000 个寓言的总成本为 USD $ 405.76，相当于每 1,000 个寓言约 USD $ 0.1353 [36] 。

这种统一的结构确保每个样本都包含叙述的精髓（`fable` ）以及复制或分析其生成所需的元数据。数据集的提示工程设计强制包含一致的元素——主要角色、特质、场景、冲突、解决方案和寓意——确保故事遵循经典寓言结构，同时保持丰富的主题多样性。因此，TF1-EN-3M 的结构性特征直接揭示了计算成本、模型行为和大规模文本输出。我们根据数据集指南 [37] 的原则设计了生成元数据架构，确保每个样本都携带再现性和伦理审查所需的所有来源、配置和评估细节。

该数据集总计约有 10 亿个模型生成文本的标记。尽管是合成的，这些故事还是相当多样化。我们进行了基本分析：

- 多样性：由于我们的提示设计，每个角色和道德的出现频率大致均匀。没有单一模板占主导地位。这种有意的分布与许多人类故事数据集中可能重复流行故事的情况形成对比。
- 质量分布：使用相同的 GPT-o3-mini 评论器，我们对最终数据集的随机样本进行了抽查。平均质量保持高水准（因为这些内容都是由所选模型生成的）。我们观察到一些如偶尔使用古语言或稍显刻意的道德教训等小问题，但这些都属于寓言文本的魅力所在，并不普遍。
- 长度：故事的平均长度为 ~ 120 个标记（不包括寓意），标准差为 ~ 30。我们指示模型生成 “短寓言”，因此没有一个故事过长。这种长度上的统一性对于训练目的很有用（较少的差异）。

数据格式：TF1-EN-3M 存储为 Hugging Face Dataset（`datasets` 库），每个条目包含提示字段和故事文本。我们在开放许可下发布它，因为它完全是机器生成的。用户可以轻松加载它用于模型训练或分析。我们强调，尽管数据是合成的，写作风格旨在类似于人类书写的寓言，且初步的人类阅读发现这些故事是合乎逻辑并且令人愉悦的。

可用性：该数据集在 Hugging Face Hub 上以标识符 `klusai/ds-tf1-en-3m` [12] 发布。研究人员和从业者可以完整下载或部分下载（为了方便，数据集被分块）。此外，我们提供了 TinyFabulist GitHub 仓库 [38] ，其中包含重新生成数据集的代码，包括：

- 每个元素的提示列表（这样就可以修改或扩展它们）。
- 使用的生成脚本（用于 Hugging Face `transformers` 或 `peft` 库等，带有我们的模型权重或对其的引用）。
- 评估脚本（用于 GPT-o3-mini 评分和翻译测试）。
- 关于如何再现该过程或创建多语言版本的指南（例如，更换为法语翻译的道德列表以获得法语寓言数据集，这是一个计划中的扩展）。

我们相信，释放这些资源将能够实现完全的可重复性，并鼓励其他人以 TinyFabulist 为基础进行构建。TF1-EN-3M 的潜在用途包括：微调较小的模型以作为寓言生成器或道德推理评估器，使用这些故事训练分类器或用于道德内容的问答模型，甚至作为计算语言学中文学研究的创意语料库 [1] 。

5 讨论与有效性的威胁

TF1-EN-3M 合成寓言数据集为进一步探索开辟了几条途径。在这里，我们讨论我们工作的更广泛影响、方法学发现及潜在应用，特别是在 1 节中提出的研究问题背景下。

我们的第一个研究问题是，组合提示扩展方法能否有效利用大型语言模型生成多样化和高质量的寓言。在 300 万篇生成的故事中，我们发现将结构化模板与六个受控输入领域（角色、特征、背景、冲突、解决、道德）的均匀采样结合，可以在保持连贯性的同时产生高度的叙事多样性。GPT-o3-mini 对语法、创造力、道德清晰度和提示遵循性的评估显示，即使是中小型模型，在提供足够的提示支架时，也可以可靠地生成内容丰富、结构良好的道德故事。此外，我们的模板设计的灵活性确保涵盖广泛的主题空间，而我们的过滤和平衡程序有助于避免模式崩溃或刻板情景的过度表示。

www.xueshuxiangzi.com

我们的第二个研究问题侧重于在资源限制下识别表现最佳的开源权重 LLM。通过对十个公开可用的指令调优模型（参数范围从 1B 到 8B）进行控制评估，我们发现包括 `Falcon3-7B-Instruct` [29] 、`Llama-3.1-8B-instruct` [24] 和 `Mistral-7B-Instruct-v0.3` [25] 在内的多个模型在四个评价轴上始终取得了最高的平均分。值得注意的是，这些模型在某些情况下甚至胜过了更大的模型，这表明指令调优质量而不仅仅是规模是寓言生成性能的关键决定因素。评估还指出了一个有利的权衡：较小的模型如 `Phi-3-mini-4k-instruct` [28] 和 `SmolLM2-1.7B-Instruct` [21] 表现出强大的语法和道德清晰度，同时推理速度较快，使它们非常适合低延迟应用和设备上的部署。最终，我们选择了 `Llama-3.1-8B-instruct` 作为总体上表现最佳的模型。

效率和可访问性。 通过证明故事生成可以通过相对较小的模型实现，我们的工作强调了自然语言生成中效率的重要性 [35] 。并不是所有的应用或社区都能负担得起部署如 GPT-4 这样的巨大语言模型所需的计算资源 [39] 。像 TF1-EN-3M 这样的语料库可以帮助启动较小的模型用于创意写作任务，从而扩大访问范围。例如，教育工作者或独立游戏开发者可以在具有普通硬件的情况下对一个 6B 或 1.3B 参数模型在 TF1-EN-3M 上进行微调，以生成道德故事或任务叙述，这种方法类似于 TinyStories [4] 。

道德和教育 AI 应用 寓言长期以来被用于传递价值观和社会规范，使其成为 AI 驱动教育工具的一个引人注目的载体 [1] 。一个辅导系统可以向学生展示一个动态生成的寓言，随后提出关于道德教训的理解和反思问题。由于每个 TF1-EN-3M 条目都明确编码了一个道德，因此可以训练模型将故事映射到道德，或检测给定叙述是否包含道德教训——这可能为 AI 的审核或生成检查系统提供信息。

尽管结果总体上是积极的，但仍需认识到合成寓言的局限性。它们往往遵循源自提示结构的陈旧模板，以动物对话和童话故事为特色。这并不能涵盖当代伦理困境的复杂性。仅仅接受 TF1-EN-3M 训练的模型可能缺乏处理细微或现代道德问题的复杂性。未来的工作可以扩展 TF1-EN-3M，包括具有模糊或多层次道德的寓言，从而增强该数据集在复杂伦理推理建模中的实用性。

将 TF1-EN-3M 与人工编写的语料库结合起来，可以产生更丰富的风格和主题多样性。将合成的寓言与人工撰写的故事混合，可以在模型驱动生成的一致性与人类散文的创造力之间取得平衡。此外，利用 TF1-EN-3M 训练的模型可以通过诸如故事填空测试的基准进行评估，从而提供关于其生成和理解具有伦理影响的连贯故事结尾能力的洞察。

在我们的流程中，GPT-o3-mini 主要作为离线评估器和批评者。未来的系统可能会整合模型反馈循环，其中大型语言模型动态地修订或批评生成过程中的寓言，这可能提高效率和质量。然而，这将需要额外的计算开销以及对批评者-生成器交互的仔细控制。

我们提出，TF1-EN-3M 可以作为评估生成模型中道德推理的基准。可以基于 TF1-EN-3M 构建诸如道德推理（根据故事预测正确的道德）或道德生成（生成符合给定道德的故事）等任务，从而促进对叙事对齐、常识推理和教学文本生成的更广泛研究。

5.1 有效性威胁

尽管 TF1-EN-3M 数据集和附带的方法论表现出令人鼓舞的结果，但在评估我们研究结果的稳健性和普遍性时，应该承认几个有效性威胁。

5.2 结构效度

一个主要关注点在于依赖于基于 LLM 的评估——特别是 GPT-o3-mini——来评估诸如道德明确性、创造力和连贯性等属性。尽管最近的研究表明，LLM-作为评判者的范式在开放性任务中通常与人类偏好一致 [7, 8] ，但这类评估仍然是代理，可能无法完美反映人类判断，尤其是在涉及微妙叙述质量时。而且，我们评估标准中使用的标准——虽然受到教育和文学标准的启发——在操作上通过数值实现，这可能会掩盖定性的细微差别 [41] 。例如，在 1 到 10 的等级上评估寓言的“道德明确性”可能会忽略模糊但在教学上有价值的叙述。结合人类评估或众包注释进行的 LLM 评估，如之前的工作中所做的那样 [15] ，可以提供更强的结构效度。

5.3 外部效度

我们的数据集基于提示元素和主要来自西方寓言传统（如伊索 [2] ）的道德教训，这有可能在生成的故事中引入文化偏见。虽然结构化的模板允许广泛的组合变化，但生成的叙述仍可能反映出隐含的西方道德框架，从而限制跨文化背景的普遍适用性。这个问题在道德故事中尤为突出，因为价值观可以有很大不同 [42] 。未来的工作应考虑结合多样的哲学和宗教传统中的道德原则，并为 TF1 框架的多语言或文化本地化变体调整提示。

www.xueshuxiangzi.com

5.4 结论有效性

我们关于模型性能和数据集质量的结论主要来源于一个基于 LLM 的评估者——GPT-o3-mini。虽然选择该模型是为了平衡可访问性和推理能力，但依赖于单一评论员会引入评分中的模型偏差风险。此前的研究表明，不同的 LLM 在充当裁判时往往会产生显著不同的偏好或评估 [43] 。

此外，我们研究中包含的十个模型是在异构设置下训练的——在预训练语料库、参数规模、微调目标和底层架构上各不相同——这本质上会影响它们生成文本的质量和风格。这些差异强调了为什么模型输出在连贯性、创造性、道德清晰度和遵从性方面可能会有所不同。

为了减轻这些偏差和方差的来源，我们还采用了补充的无参考指标——Self-BLEU [9] 、Distinct-n [10] 和 Flesch 阅读容易度 [11] ——捕捉多样性、词汇丰富性和可读性，这些不需要人类或大型语言模型的参考。结合多种评估者和多样的指标可以增强评估的稳健性，并提供一个更细致、多方面的模型和生成故事质量的视角。

6 结论

我们介绍了 TF1-EN-3M 合成寓言数据集，该数据集是一个通过指令调优的、紧凑的、开放许可的语言模型生成的大规模道德导向短篇故事合集。我们的研究结果表明，即使是中等规模的 LLM——而不是百亿参数的庞然大物——也可以通过专注的提示工程和精心策划的生成管道生成多样化、具有伦理主题的叙事 [4] 。TF1-EN-3M 结合了合成数据增强 [30] 、故事生成和道德自然语言处理（NLP）[1] 的技术，创造了一种新颖的资源，既可以用于训练又可以用于在需要道德一致性以及语言流畅性的叙事任务上评估模型。

定量和定性评估表明，这些合成寓言展现出强烈的连贯性和道德清晰性。我们预期 TF1-EN-3M 将作为微调较小模型进行寓言生成的平台，以及研究语言模型如何学习和表示道德概念。展望未来，关键扩展可能包括扩大道德和场景的广度，探索专门用于叙事创作的架构，或结合人类参与的反馈以进一步提升质量。

在更广泛的人工智能领域中，该项目与发展文化和道德敏感的系统的目标一致。通过使较小、更易接触的模型来生成充满价值观的故事，我们向着不仅技术高效，而且在社会上更有根基的人工智能更进一步。我们邀请研究界利用并发展 TF1-EN-3M，相信这将推动低资源故事生成、道德内容创作，以及在语言建模中整合道德推理方面的进步。

References

[1] Jian Guan, Ziqi Liu, and Minlie Huang. A Corpus for Understanding and Generating Moral Stories, April 2022. arXiv:2204.09438 [cs].

[2] Aesop. *Aesop's Fables*. OUP Oxford, July 2002. Google-Books-ID: n2LlrCeYl7gC.

[3] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey, June 2024. arXiv:2406.15126 [cs].

[4] Ronen Eldan and Yuanzhi Li. TinyStories: How Small Can Language Models Be and Still Speak Coherent English?, May 2023. arXiv:2305.07759 [cs].

[5] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks Are All You Need, October 2023. arXiv:2306.11644 [cs].

[6] Mihai Nadas, Laura Diosan, and Andreea Tomescu. Synthetic Data Generation Using Large Language Models: Advances in Text and Code, March 2025. arXiv:2503.14023 [cs].

[7] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment, May 2023. arXiv:2303.16634 [cs].

[8] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as You Desire, February 2023. arXiv:2302.04166 [cs].

[9] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A Benchmarking Platform for Text Generation Models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1097–1100, New York, NY, USA, June 2018. Association for Computing Machinery.

[10] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A Diversity-Promoting Objective Function for Neural Conversation Models, June 2016. arXiv:1510.03055 [cs].

www.xueshuxiangzi.com

[11] J. P. Kincaid and And Others. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, National Technical Information Service, Springfield, Virginia 22151 (AD-A006 655/5GA, MF $ 2, February 1975. ERIC Number: ED108134.

[12] klusai/ds-tf1-en-3m · Datasets at Hugging Face.

[13] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What Makes Good In-Context Examples for GPT- $ 3 $?, January 2021. arXiv:2101.06804 [cs].

[14] Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. Reframing Instructional Prompts to GPTk's Language, March 2022. arXiv:2109.07830 [cs].

[15] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical Neural Story Generation, May 2018. arXiv:1805.04833 [cs].

[16] Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-and-Write: Towards Better Automatic Storytelling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7378–7385, July 2019. Number: 01.

[17] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada, July 2023. Association for Computational Linguistics.

[18] Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization, October 2023. arXiv:2305.13091 [cs].

[19] Angela Naimou. Short Fiction, Flash Fiction, Microfiction. In Joshua Miller, editor, *The Cambridge Companion to Twenty-First Century American Fiction*, Cambridge Companions to Literature, pages 21–42. Cambridge University Press, Cambridge, 2021.

[20] Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. Trading Off Diversity and Quality in Natural Language Generation. *ArXiv*, April 2020.

[21] HuggingFaceTB/SmolLM2-1.7B-Instruct · Hugging Face, February 2025.

[22] CohereForAI/aya-23-8B · Hugging Face, March 2025.

[23] meta-llama/Llama-3.2-1B-Instruct · Hugging Face, December 2024.

[24] meta-llama/Llama-3.1-8B-Instruct · Hugging Face, December 2024.

[25] mistralai/Mistral-7B-Instruct-v0.3 · Hugging Face.

[26] Qwen/Qwen2.5-7B-Instruct · Hugging Face, February 2025.

[27] deepseek-ai/deepseek-llm-7b-chat · Hugging Face, August 2024.

[28] microsoft/Phi-3-mini-4k-instruct · Hugging Face, January 2025.

[29] tiiuae/Falcon3-7B-Instruct · Hugging Face, February 2025.

[30] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations, October 2023. arXiv:2310.07849 [cs].

[31] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling Synthetic Data Creation with 1,000,000,000 Personas, September 2024. arXiv:2406.20094 [cs].

[32] 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 Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada, July 2023. Association for Computational Linguistics.

[33] M. O. Riedl and R. M. Young. Narrative Planning: Balancing Plot and Character. *Journal of Artificial Intelligence Research*, 39:217–268, September 2010.

[34] Huyen Nguyen, Haihua Chen, Lavanya Pobbathi, and Junhua Ding. A Comparative Study of Quality Evaluation Methods for Text Summarization, June 2024. arXiv:2407.00747 [cs].

[35] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, January 2020. arXiv:2001.08361 [cs].

www.xueshuxiangzi.com

[36] Pricing.

[37] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, December 2021.

[38] klusai/tinyfabulist, April 2025. original-date: 2025-01-07T20:20:42Z.

[39] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, ukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, ukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report, March 2024. arXiv:2303.08774 [cs].

[40] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California, June 2016. Association for Computational Linguistics.

[41] Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A. Raffel. Scaling Data-Constrained Language Models. *Advances in Neural Information Processing Systems*, 36:50358–50376, December 2023.

[42] Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences, December 2020. arXiv:2012.15738 [cs].

www.xueshuxiangzi.com

[43] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, December 2023. arXiv:2306.05685 [cs].

[44] NVIDIA L40S GPU.

[45] AI Chip - AWS Inferentia - AWS.

[46] NVIDIA A100 GPUs Power the Modern Data Center.

[47] NVIDIA L4 Tensor Core GPU.

A 硬件和环境配置

我们在多个 GPU 配置下使用 `Llama3.18BInstruct` 对 TF1EN3M 数据集进行推理基准测试，使用相同的提示和解码设置。所有实验均在 HuggingăFace 推理端点上进行；我们使用 HuggingăFace 在 2025 年 4 月公布的小时费率计算成本。表 5 报告了生成固定大小批次的寓言的实时时间和可计费成本，而表 6 使用来自供应商文档的架构规格对这些结果进行了背景说明。

Table 5: 固定提示批次 (`Llama3.18BInstruct`) 的经验推理时间和 Hugging Face 终端成本。

Hardware (HFăinstance)	Timestamp Range	HFăRate (USD/h)	Duration (min)	Cost (USD)
L40S	20250412 22:15ă–ă22:30	$ 1.80	15.4	$ 0.46
A10G	20250412 22:15ă–ă23:16	$ 1.00	61.1	$ 1.02
A100	20250412 22:15ă–ă22:28	$ 4.00	12.7	$ 0.85
L4	20250412 22:15ă–ă00:06	$ 0.80	110.9	$ 1.48

推理时间列记录了生成日志中最早和最晚时间戳之间的间隔。成本计算为 Rate $\times \frac{\text{time (s)}}{3600}$ 。Hugging Face 按分钟计费，向上舍入到下一个分钟 [1] 。所有任务都使用 Hugging Face Text Generation Inference (TGI) 作为后端。

尽管 A100 提供了最快的处理速度，但其较高的费用缩小了价格差距：对于我们的批量大小，L40S 实现了最佳的时间成本平衡，这与 NVIDIA 将 L40S 定位为高吞吐量 GenAI 推理的说法一致。

Table 6: AWS Inferentia 2 和具有代表性的 NVIDIA GPU 的规格比较。小时价格对应于 Hugging Face 推理端点的列表费率（2025 年 4 月）。

Feature	Inferentiaă2	A10G	A100	L4	L40S	T4
Type	CustomăAWSăsilicon	DCăGPU	DCăGPU	DCăGPU	WS/DCăGPU	DCăGPU
ReleaseăYear	2023	2021	2020	2023	2023	2018
Architecture	NeuronCorev2	AmpereăGA102	AmpereăGA100	AdaăLovelace	AdaăLovelace	TuringăTU104
GPUăMemory	N/A (SRAM)	24ăGBăGDDR6	40/80ăGBăHBM2e	24ăGBăGDDR6	48ăGBăGDDR6	16ăGBăGDDR6
INT8ăTFLOPS	∼ 400	312	624/312	1ă466/733	2ă805/1ă402	260
FP16ăTFLOPS	∼ 100	124	312	183	742	65
BF16ăSupport	Yes	No	Yes	Yes	Yes	No
InferenceOptimised	Yes	Moderate	Train & largeăinf.	Yes	Balanced	Yes
Poweră(W)	∼ 150	150–300	400	72	300–350	70
FormăFactor	AWSăonly	PCIe	SXMă/ăPCIe	PCIe	PCIe	PCIe
CloudăAvailability	AWSăonly	AWS/GCP/Azure	AWS/GCP/Azure	GCP/Azure	AWSă(roadmap)	AWS/GCP/Azure
TypicalăUseăCases	Highthroughputăinf.	Balancedăinf.	Traină & ălargeăinf.	Realtimeăinf.	GenAIă/ăvisual	Costeff.ăinf.
HFăEndpointă $ /hr	$ 1.20	$ 1.00	$ 4.00	$ 0.80	$ 1.80	$ 0.60

供应商数据显示，AWS Inferentia 2 专为大规模、低延迟的推理而设计，而 NVIDIA 的产品组合涵盖了面向成本的 (T4)、平衡的 (A10G) 和旗舰 (A100) 加速器。NVIDIA L4 在边缘和对延迟敏感的部署中提供了每瓦的卓越性能，而更新的 L40S 则以 FP8 和第四代张量核为特点，专注于高端生成工作负载。将这些公开的功能与我们的实测时间（表 5 ）结合起来，厘清了生产数百万合成寓言的成本性能界限：在我们的情况下，L40S 实现了单个寓言的最低成本，但对延迟要求极低的工作负载可能仍然需要支付 A100 或 Inferentia 2 的溢价。

[1] 参见 HF 定价页面 [36] 。

www.xueshuxiangzi.com