Grokking in the Wild: Data Augmentation for Real-World Multi-Hop Reasoning with Transformers

Roman Abramov [1] Felix Steinbauer [1] Gjergji Kasneci [1 2]

英寸

Abstract

Transformer 在众多 NLP 任务上取得了很大的成功，但在多步骤事实推理中仍表现出显著差距，尤其是在现实世界知识稀疏的情况下。最近的 groking 研究表明，神经网络一旦发现潜在的逻辑模式，就能从记忆过渡到完美泛化——然而这些研究主要使用小型合成任务。在本文中，我们首次将 groking 扩展到现实世界的事实数据，通过用精心设计的合成数据增强现有知识图谱来解决数据集稀疏性的问题，以将推断事实与原子事实的比例提高到 ϕ_r 超过 groking 所需的门槛。令人惊讶的是，我们发现即使事实不正确的合成数据也可以强化生成的推理电路，而不是降低准确性，因为它迫使模型依赖关系结构而不是记忆。在多跳推理基准测试中，我们的办法在 2WikiMultiHopQA 上取得了最多 95–100 % 的准确率——显著超越了强基线，并匹配或超过当前先进的结果。我们还深入分析了 ϕ_r 增加如何驱动 Transformer 内部的泛化电路的形成。我们的研究结果表明，基于 groking 的数据增强可以开发隐式多跳推理能力，为在大规模语言模型中实现更为稳健和可解释的事实推理打开了大门。

1 介绍和相关工作

变压器在文本分类、摘要和机器翻译等广泛的自然语言处理（NLP）任务中表现出了显著的成功。然而，当要求它们执行多步或多跳的事实推理时，尤其是在知识既庞大又分散的真实场景中，它们仍然面临重大挑战。造成这种困难的一个关键原因在于模型倾向于记忆而不是泛化——这个问题在知识密集型任务中由于数据分布不够丰富而变得尤为严重。

最近关于 grokking 的研究表明，在某些条件下，过参数化的神经网络在长时间训练后，会突然从纯粹的记忆转变为近乎完美的泛化。早期研究通常集中于高度控制的合成任务，如模运算或简化的算法数据集。在这些玩具设置中，推断事实的数量（即多步或复合模式）可以系统地增加，直到达到一个临界比率 ϕ ，此时模型中会出现一个 “泛化电路”。

然而，现实世界的数据集形成鲜明对比：事实性知识极为稀疏，且常常分散在不完整或嘈杂的知识图谱中。因此，核心挑战是确保相对于原子事实（直接陈述）而言，有足够的高阶推断事实，以支持 grokking 所需的内部电路形成过程。换句话说，在现实世界的场景中，人们不能简单地保证多步（推断）事实与单跳（原子）事实之间有足够大的比率 ϕ 。我们的工作正是解决这一障碍，通过提出一种数据合成策略，以增强和重新平衡现实世界的知识库。

多跳问答 (QA)。 2WikiMultiHopQA (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022) 特别适合于评估多步骤的事实推理。这个数据集包含基于维基百科的查询，需要在给出答案之前检索和组合多个证据（分布在不同页面或段落中）。这一特性与我们对多跳推理的关注紧密相连，并强调了隐式推理能力的必要性：在许多情况下，系统必须链接和遍历几个事实节点——例如，“米歇尔是奥巴马的妻子” 和 “米歇尔出生于 1964 年”——以得出最终结论（例如，关于米歇尔的出生年份或其他推导出的事实）。此外，这些体现了一个大的、复杂的知识图谱，具有真实世界的实体、模糊的语言和长尾关系——所有这些都使它们成为大规模事实推理的典型测试床。

洞察及其在 Transformer 泛化中的作用。 Grokking 的概念在 Power et al. (2022) 中被引入，展示了神经网络不仅可以学习表面的模式，还可以在长时间训练中通过合适的归纳偏置学习更深层次的、可推广的推理机制。随后工作的研究将 grokking 与双重下降 (Belkin et al., 2019; Nakkiran et al., 2020) 、深网络损失景观的几何 (Davies et al., 2023) 和权重衰减 (Pezeshki et al., 2022; Nanda et al., 2023) 相关联，表明适当的正则化可以鼓励通用电路的出现。这些电路一旦形成，能够进行超出简单记忆的分布外推理 (Varma et al., 2023; Liu et al., 2023) 。

文献中的空白。 尽管在知识图谱补全 (Liu et al., 2022) 和通过基于检索的方法进行多跳问题回答 (Yang et al., 2018; Ho et al., 2020) 方面进行了广泛的研究，但很少有研究考察是否可以在真实世界文本环境中利用内部 grokking 现象用于隐式多跳推理。大多数先前的方法要么：

[1] 慕尼黑工业大学计算、信息与技术学院，德国慕尼黑 [2] 慕尼黑工业大学社会科学与技术学院，德国慕尼黑. Correspondence to: Roman Abramov < roman.abramov@tum.de >, Felix Steinbauer < felix.steinbauer@tum.de >.

- 专注于玩具任务：例如，模数加法或合成数学问题 (Power et al., 2022; Nanda et al., 2023; Wang et al., 2024) 。
- 依靠显式提示或连锁思维：其中中间推理步骤必须在外部文本 (Plaat et al., 2024) 中被明确指出，而不是作为隐含电路来学习。
- 使用标准的图补全架构：例如，基于 GNN 的解决方案来扩充部分知识图 (Liu et al., 2022) ，这不一定会导致后期的内部电路形成或突发的泛化。

据我们所知，还没有现有的研究使用基于 grokking 的方法来展示 Transformer 如何在大规模事实数据上隐式地发现多跳推理能力。

贡献。 在本文中，我们通过以下贡献来弥补这一空白：

- 我们采用了一种针对性的数据合成程序，以确保每个关系的 ϕ 足够大，从而释放在真实世界中基于维基百科任务形成内部泛化电路的潜力。
- 我们表明，即使是事实不正确的合成数据，也可以提高推测事实与原子事实的比例，通常是增强而不是损害逻辑一致性。
- 我们在 2WikiMultiHopQA 上的实验表明，一旦 ϕ 的比例超过某个阈值，顿悟就会出现——使得 Transformer 能够在没有显式思维链提示或复杂外部支撑的情况下执行复杂的多步推理。

在以下章节中，我们详述了数据增强策略的数学基础、2WikiMultiHopQA 上的实证设置，以及关于隐式多跳推理的新见解。我们的研究发现表明，grokking 不是限制在人为构造的数据集上的产物，而是一种强大的机制，通过适当的数据分布调整，可以被用于大规模的真实世界事实推理。

2 问题描述

多跳推理的概念预设了一个知识图谱（KG），其节点（实体）和边（关系）可以通过一系列推理步骤进行遍历。在我们的设定中，这个 KG 被编码为文本形式，但在结构上，我们仍然在处理基于 KG 的多跳问答（Multi-hop KGQA）。先前的研究 (Liu et al., 2022) 已经观察到，完成知识图谱对于多跳 KGQA 来说可能是至关重要的；为了形成基于 grokking 的 Transformer 泛化电路，扩充原始 KG 以便存在足够多的多步（推断）事实是必需的。本节正式化了我们为了实现 Transformer grokking 而增强 KG 的问题。

2.1 定义和基础

一开始，我们引入关键符号以便清晰理解。读者可以随时参考表 1 ，查看主要符号的简要总结。

知识图谱。

Definition 2.1 (Knowledge Graph). 我们将知识图定义为一个元组 $\mathcal{KG} = (\mathcal{V}, \mathcal{R}, \mathcal{F}_A)$, ，其中

- $\mathcal{V}$ 是一组有限的实体（节点或顶点），
- $\mathcal{R}$ 是一组有限的关系类型（边/谓词），
- $\mathcal{E} \equiv \mathcal{F}_A \subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V}$ 是一个有限的原子事实（三元组）集合，形式为 (h, r, t) ，其中 $h \in \mathcal{V}$ 是头部（主语），$t \in \mathcal{V}$ 是尾部（宾语），$r \in \mathcal{R}$ 是关系。

在自然语言中，实体通常由关系陈述连接。虽然一些关系（例如，“之子”）是有向的，但由于一个陈述通常可以双向查询（主语/宾语），人们也可以将 KG 视为无向进行图遍历。

Definition 2.2 (Norm over edges). 对于在 $\mathcal{KG}$ 中，将严格定向的边视为无向边进行计数，我们使用简单的计数：$|\,| = \sum_{(h,\ldots,t)\in\mathcal{E}} 1$.

Definition 2.3 (Average branching factor). 一个知识图谱的平均分支因子是：$b = \frac{|\mathcal{F}_A|}{|\mathcal{V}|}$ 。

Definition 2.4 (Atomic Facts). 遵循先前的工作，我们使用 $\mathcal{F}_A$ （或同等地 $\mathcal{F}_1$ ）来表示原子事实的集合，即在知识图中显式存储的所有一阶三元组。

与显式推理方法（例如，链式思维提示 (Plaat et al., 2024) ）不同，隐式推理依赖于模型在训练过程中形成内部电路 (Nanda et al., 2023) 。“Grokking” 研究 (Power et al., 2022; Wang et al., 2024) 表明，在某些条件下，Transformers 能够通过学习这些内部电路，从记忆转向完全的泛化。

如 Wang et al. (2024) 所指出的，对于电路形成的关键因素是每个关系中推导事实与原子事实的比例 r 。设

$$\mathcal{F}_{A,r} \subseteq \mathcal{F}_A \quad \text{and} \quad \mathcal{F}_{I,r} \subseteq \mathcal{F}_I$$

为涉及关系 r 的原子和推导事实的集合。那么：

当 ϕ_r 超过某个阈值时——实证发现对于 GPT-2 风格 Transformers 而言，这个阈值为慢速泛化约为 3.6，快速电路则高达 18——模型倾向于为该关系形成一个泛化电路 (Power et al., 2022) 。这些阈值是近似的且依赖于架构 (Nanda et al., 2023; Wang et al., 2024) ，但说明了核心原则：如果没有足够多的多跳事实，模型就永远无法“掌握”底层关系。

即使一个知识库只在某个方面富含多跳数据（即，一些关系满足阈值 ϕ_G ，而其他关系不满足），它仍然能对某些推理任务带来好处。然而，要实现完全的泛化性，所有关系必须超过该阈值：

Definition 2.5 (Generalizable Knowledge Base). 如果知识库 $\mathcal{KG}$ 可以满足

$$\forall r \in \mathcal{R}\colon \phi_r \geq \phi_G,$$

，其中 ϕ_G 是给定模型设置所需的最小泛化比率，那么我们称其为可泛化。如果这一条件仅适用于关系的一个子集，我们称之为 $\mathcal{KG}$ 部分可泛化。

由于每个关系 r 都有自己的比例 ϕ_r ，如果任何 ϕ_r 过低，知识库（KB）可能无法支持掌握。分析界限（见附录 A.3 ）揭示：

- 可以通过添加新节点或增加与 r 相关的边来增加比例 ϕ_r ，但这种效果受到图的连接方式以及关系 r 出现频率的限制。

- 典型的现实世界知识图谱往往是稀疏的，这就是为什么数据合成（在后面章节将描述）对于提升 ϕ_r 至关重要的原因。

Example 1 (Insufficient Branching Factor). 假设一个家族 KG 有 $\{\text{father}, \text{mother}, \text{child}_1, \ldots\}$ ，边像“父母”，“兄弟姐妹”，“拥有者”。虽然从表面上看图是很好地连接着，其特定关系的分支因子可能仍然太小而无法通过阈值 $\phi_G \approx 3.6$ 。因此，KB 仍然无法完全泛化，阻止 Transformer 为所有关系形成强大的多跳电路。

2.2 设置和查询

在实际应用中，我们通过提出问题（其唯一答案位于简单的（无环）推理路径的末端）来测试模型的多跳推理能力。例如：

Example 2 (Chaining 3 Steps). 文本查询“哪部电影在奥巴马夫人出生的同一年上映？”对应于一个三步推理：

$$(\text{“Obama”}, \text{“wife of”}, \text{“born in”}, \text{“aired in”}, t).$$

这里，模型必须在内部推断：

$$I(\text{“Obama”}, \text{“wife of”}) \rightarrow \text{“Michelle”},$$
$$I(\text{“Michelle”}, \text{“born in”}) \rightarrow \text{“1964”},$$
$$I(\text{“1964”}, \text{“aired in”}) \rightarrow t = \text{“Mary Poppins”}.$$

成功的隐含推理要求模型既能记住原子事实，又能通过一个通用化的电路将它们串联在一起。

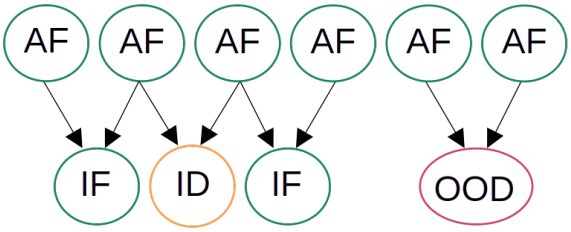

Figure 1. ID 和 OOD 的概念性差异：在训练中，展示了在分布内（ID，橙色）和分布外（OOD，红色）推导的事实。所有绿色组件在训练期间被看到，包括所有原子事实（AF）和一些推导的事实（IF）。

Definition 2.6 (In-distribution vs. Out-of-distribution).
分布内 (ID)：从训练数据中存在的原子事实的组合中推导出的事实，这些事实从未以这种特定组合出现在一起。模型单独见过所有的知识成分，并且涉及它们的不同推理组合，但在这个特定的测试安排中没有见过。
分布外 (OOD)：从训练数据中存在的原子事实中推导出的事实，这些事实从未用于任何训练推理路径中。模型见过单个的知识成分，但不知道在被测试的背景中它们应该如何应用。
图 1 显示了与 ID 和 OOD 测试事实相关的已训练事实（原子和推导）。

Table 1. 关键符号表。本论文中经常使用的符号。

Symbol	Meaning
$\mathcal{V}$	Set of entities (nodes) in the KG
$\mathcal{R}$	Set of relation types (edges)
$\mathcal{F}_A$	Atomic facts (1-hop) $\subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V}$
$\mathcal{F}_I$	Inferred facts (multi-hop) $\bigcup_{n \geq 2} \mathcal{F}_n$
b	Average branching factor $\frac{\lvert\mathcal{F}_A\rvert}{\lvert\mathcal{V}\rvert}$
ϕ_r	$\frac{\lvert\mathcal{F}_{I,r}\rvert}{\lvert\mathcal{F}_{A,r}\rvert}$ (ratio of inferred to atomic facts)
ϕ_G	Minimal generalization ratio for successful grokking
$I(h, r)$	Inference step: from entity h via relation r to a neighbor

正如我们在实验中所展示的，OOD 查询特别具有挑战性，需要真正的结构化泛化，而不是部分记忆。这正是充分高的 ϕ_r 比率对于启用 Transformer 从局部记忆向稳健的多跳推理过渡变得关键的地方。

2.3 引理与界限

我们介绍的第一个引理明确指出，（1）选择节点与在路径中排列它们与边缘存在性之间会导致渐进的上限，以及（2）即使是大型知识库（KB）如果没有额外的数据增强，也不会超过大约 b^{n-1} 增长 $\phi_{n,r}$ 。

引理 1（关于 n 跳路径数量的渐近界）。 考虑一个知识库（KB），具有 $|\mathcal{V}|$ 实体，平均分支因子 b ，以及一个（近似）随机图假设，即两个不同节点之间的每个潜在 directed 边存在的概率为 $\frac{b}{|\mathcal{V}|-1}$ 。那么，对于较大的 $|\mathcal{V}|$ ，有效的 n 跳路径的期望数量满足：

$$|\mathcal{F}_n| \approx \binom{|\mathcal{V}|}{n+1} (n+1)! \left(\frac{b}{|\mathcal{V}|-1}\right)^n.$$

此外，在限制 $|\mathcal{V}| \to \infty$ 的情况下，关系特定的比率 $\phi_{n,r}$ （即关系 r 的 n 跳到 1 跳事实的比例）保持在 b^{n-1} 之上。关于引理 1 的证明，请参见附录 A.1 。关于节点数量 $|\mathcal{V}|$ 的较低泛化界限，请参见附录 A.3 。第二个引理明确指出，每个关系的分支因子 b_r 必须足够大以超越经验门槛 ϕ_G 。如果一个或多个关系低于该门槛，就不会发生全关系的完全泛化——尽管对于较高 b_r 的关系可能会出现部分泛化。

引理 2（完全可推广性的必要条件）。 设 ϕ_G 是触发基于 grokking 的泛化所需的最小比例，针对给定的模型架构。一个知识库只有在 n 跳事实上完全可泛化时才满足 $\forall r \in \mathcal{R}$

$$\phi_{n,r} \geq \phi_G \quad \Rightarrow \quad b_r > \sqrt[n-1]{\frac{\phi_G |\mathcal{V}| (|\mathcal{V}|-1)^n}{\binom{|\mathcal{V}|}{n+1} (n+1)!}},$$

其中 $b_r = \frac{|\mathcal{F}_{A,r}|}{|\mathcal{V}|}$ 是关系特定的分支因子。相同地，如果任何 b_r 低于该门槛，则知识库不能完全可泛化。

关于引理 2 的证明草图，请参见附录 A.2 。

3 方法

我们的方法遵循一个两阶段的流程，旨在实现实际世界中多跳推理任务的掌握。首先，我们通过合成知

识扩充原始 Wiki2Hop 数据集，提高多跳（推断）事实与单跳（原子）事实的比例。其次，我们训练一个 Transformer 模型数十万步，利用掌握特有的后期泛化现象 (Power et al., 2022; Wang et al., 2024) 。

3.1 数据集

2WikiMultiHopQA 概述。 我们使用 2WikiMultiHopQA 数据集 (Ho et al., 2020) ，这是一个知名的检索增强生成（RAG）和多跳 QA 基准。2WikiMultiHopQA 由维基百科段落、支持事实（三元组）和通常需要链接多条证据的推理查询组成。尽管其范围很广，但初始比例 $\phi \approx 0.5$ ——即每两个原子事实仅产生一个推断事实——使得其不足以让掌握自然地出现。

结构化与非结构化。 我们将 2WikiMultiHopQA 分为两个子集：

- 结构化：支持事实被简化为简短的三元组（例如，Paris – country – France ）。
- 非结构化的：支持性的事实嵌入在完整的维基百科段落中，提供了更丰富的背景，但也带来了更多的噪音和复杂性。

在这些子集内，我们专注于四个主要多跳任务中的两个：

1. 比较，比较不同实体的属性（例如，相同的位置或发行年份），
2. 组合，通过链接多重关系来推导出答案（例如，谁执导了某部电影的续集?）。

我们的目标是提升 ϕ 以便比较和组成查询，这样 Transformer 可以发展内部推理回路。

3.2 数据增强

为了提高多跳问题的覆盖率，我们通过基于 LLM 的生成系统地添加原子和推断事实。在每种情况下，我们确保与 2WikiMultiHopQA 风格一致，保持类分布平衡，并密切跟踪 ϕ ，使其超过已知的掌握阈值 (Wang et al., 2024) 。

3.2.1 比较任务

在比较任务中，原子事实（例如，City – country – X ）被配对形成关于共同属性的问题。起初，我们选择了 120 个原子事实和 60 个推理事实，重点关注地理位置（印度、法国、美国、加拿大、俄罗斯）。利用我们的增强策略，我们将其扩展为 1,000 个原子事实和 8,000 个推理事实，得到了 $\phi_G = 8$ ——远高于 Wang et al. (2024) 为慢理解报告的最低比例 3.6 。

生成新位置。 首先，我们生成描述原始集合中不存在的额外城市或区域的原子事实。对于结构化版本，这些保持在 (City, country, Country) 格式中。对于非结构化数据，我们提示大型语言模型（LLM）生成简洁、类似维基百科的段落（例如，“巴黎卢浮宫：卢浮宫是一座世界著名的艺术博物馆……”）。

Algorithm 1 用于比较的增强算法

```
Require: loc_examples   // sample atomic facts
Require: detailed_examples   // extended paragraphs
atomic ← generate_locations (loc_examples)
if task_type == “full_text” then
    atomic ← detalize_locations (atomic, detailed_examples)
end if
inferred ← generate_inferred (atomic)
return atomic, inferred
```

Algorithm 2 用于组合的增强算法

```
Require: text   // original atomic facts in textual form
graph ← parse_graph (text)
atomic ← graph. augment_atomic ()
inferred ← graph. augment_inferred ()
inferred ← diversify (inferred)
return atomic, inferred
```

生成推断示例。 接下来，我们通过选择两个不同的位置事实并提示，例如，“阿维尼翁的圣母岩和巴黎卢浮宫都位于同一个国家吗?”来创建基于比较的查询。算法 1 给出了该过程的伪代码，其中 generate_inferred 系统地将原子事实合并为比较问题：

3.2.2 组合任务

组合任务需要顺序链接多个关系（例如，Person – spouse of – X – born in – Year ）。我们从 200 个原子事实和 100 个推理事实开始，确保答案中不直接提及日期，并扩展至 800 个原子事实加上 5,000 个推理事实。这将 ϕ_G 提升到 6.25 足以在实践中触发 whacking 动态。

图处理。 我们将文本事实转换为图表示，通过 LLM 基础解析器提取节点（实体）和边（关系）。然后，我们通过添加新的原子边（避免循环）并采样多跳路径来丰富这个图，从而产生额外的推理事实。

推断问题五元组。 每个长度为二或三的多跳路径对应一个五元组 $(obj_1, rel_1, obj_2, rel_2, obj_3)$ 。然后，一个 LLM 将这些五元组转化为自然语言问题（例如，“为什么 Dunsany 的第 19 任男爵 Randal Plunkett 的父亲去世?”），这接近真实的 2WikiMultiHopQA 风格。算法 2 详细描述了这一方法：

增强事实示例。 表 2 提供了一个样本，展示了一个原子事实、一个详细版本以及一个推断问题在扩充后如何呈现。注意最终的问题如何优雅地将两个位置联系在一个关于它们是否属于同一国家的是/否查询中。

在这些增强步骤之后，生成的数据集达到了足够高的 ϕ 。在下一节中，我们详细说明如何在这个丰富的语料库上训练一个 GPT-2 风格的 Transformer，以及如何通过延长优化揭示多跳推理准确性在后期的显著跃升。

4 实验

我们在 2WikiMultiHopQA 数据集 (Ho et al., 2020) 上评估基于 grokking 的方法，并进行扩充以确保足够大

Table 2. 2WikiMultiHopQA 比较任务的扩展事实示例

Type	Example
Atomic Fact	Louvre Museum – country – France
Detailed Fact	Paris Louvre Museum: The Louvre Museum is a world-famous art museum...
Inferred Question	Are Avignon Rocher des Doms and Paris Louvre Museum both located in the same country? (Answer: Yes)

的比例 ϕ 。我们报告了结构化（?? ）和非结构化（?? ）子集以及原始（未扩充）数据（?? ）的结果。最后，我们比较所有设置下的性能（?? ）并提供定性的见解（?? ）。

4.1 设置

模型和训练。 我们从头开始训练一个 8 层的 GPT-2 风格的 Transformer（768 隐藏单元，12 个注意头），使用 AdamW (Loshchilov et al., 2017) ，学习率为 5×10^{-5} ，批量大小为 512，权重衰减为 1。我们依赖 HuggingFace Trainer [1] 进行默认调度，bf16 精度和 torch_compile 以提高速度。训练运行到达 ∼300k 步或在分布外（OOD）准确性出现后期跃升时停止。我们在一台 A100 GPU 上使用三个随机种子进行训练（每次 48 小时），报告最佳运行结果（变化性 ±1–2% ）。

ID 与 OOD 评估。 在分布内（ID）的查询重复使用训练中观察到的实体/关系组合，而分布外（OOD）的查询涉及全新的组合。“grokking” 的标志是在经过长时间训练后，OOD 准确性出现延迟性的显著提高（(Power et al., 2022) ）。

仅使用原始结构化比较数据训练会产生 100% 的训练准确率，没有出现后期阶段的 OOD 跃升（图 2 (a)）。这一平台期与先前的发现一致，表明 $\phi \approx 0.5$ 不足以引发 grokking (Wang et al., 2024) 。

图 2 (b) 展示了当查询询问两个实体是否共享某一属性（例如，City A – country – France vs. City B – country – France ）时，OOD 准确率的明显后期跃升。这种更简单的关系结构更容易触发泛化电路的形成。

图 2 (c) 展示了使用基于三元组的事实进行组合任务的训练曲线，形式需要 X – spouse of – Y – nationality – Z 的链条。

移至完整的 Wikipedia 段落（对比 的三元组）会增加噪声和长度（见图 2 (d)）：对于更简单的比较查询，模型仍然能够达到不错的 ID 准确率，但难以泛化到 OOD，反映出非结构化文本中有效 ϕ 的稀疏性。

表 ?? 将我们的 Grokked GPT-2–small 与 GPT-4o 和 o1-mini 在结构化数据集上进行比较。我们的方法优于其他方法，尤其是在比较任务中，即使在 OOD 环境中我们也几乎达到 100% 。预训练模型如 GPT-4o 可能无法提供明确的 ID/OOD 区分，因为它们已经看过大量的 Wikipedia 文本。

当合成增强覆盖多跳链（2–3 跳）时，模型可以处理诸如 “哪位画家与公司 X 的创始人出生在同一城市?” 或者 “城市 A 和城市 B 是否在同一国家?” 的查询——否则会在较低的 ϕ 上失败的任务。

失败案例。 罕见的关系或模糊的实体名称仍然是挑战。如果多个实体共享一个标签但数据集中缺乏适当的消歧，模型可能会混淆它们，从而阻碍 OOD 性能。附录 A.4 列举了一些例子。

总体发现。 这些结果证实，通过精心设计的合成数据增加 ϕ 是基于 grokking 的多跳 QA 的关键。虽然组合任务和非结构化文本可能需要更大或更有针对性的增强，但对于更简单的关系查询，“记忆到泛化” 的跃升表明 grokking 可以显著提升真实世界事实推理中的 OOD 性能。

5 局限性和未来工作

有多种途径可以改进和扩展我们基于领悟的多跳推理方法。虽然我们在 2WikiMultiHopQA 上的概念验证实验提供了有希望的证据，但关于数据复杂性、可解释性和资源可行性仍然存在重要的悬而未决的问题。

数据集和基准测试。 我们的结果表明，Transformer 的顿悟可以在诸如 2WikiMultiHopQA 等真实世界数据集上被引发。然而，更广泛的具有挑战性的推理基准可以阐明我们方法的真正范围和边界。例如，需要更长推理链、专业领域知识（例如生物医学）或时间推理的任务可能揭示出在标准的基于维基百科的问答中没有出现的复杂约束。

虽然我们观察到了出现的泛化电路，但这些电路形成的具体机制仍然只部分被理解。未来的工作可以：

事实性。 一个关键问题是合成数据（其中一些是故意或意外产生的幻觉）如何影响模型的事实准确性。虽然我们的实验表明适量的非事实数据可以增强泛化能力，但我们承认存在潜在风险：

- 真实世界知识的扭曲：如果不进行仔细过滤，幻觉可能会覆盖或遮蔽真实事实。
- 事实脆弱性：某些任务（例如，医学或法律推理）要求严格的正确性，任何事实上的偏离都是不可接受的。

作为部分解决方案，我们设想通过更复杂的基于约束的数据增强来保持核心事实性，同时仍然提高推断到原子比率 ϕ 。调查这种混合策略是未来研究的一个有趣方向。

我们将范围从人为设计的玩具问题扩展到从维基百科中得出的大型、真实的数据集。但是，仍然有很多待解的问题：

可行性。 最后，我们注意到，为了进行 grokking 而在长时间内训练大型 Transformer 架构可能会非常昂贵。

[1] https://github.com/huggingface/transformers

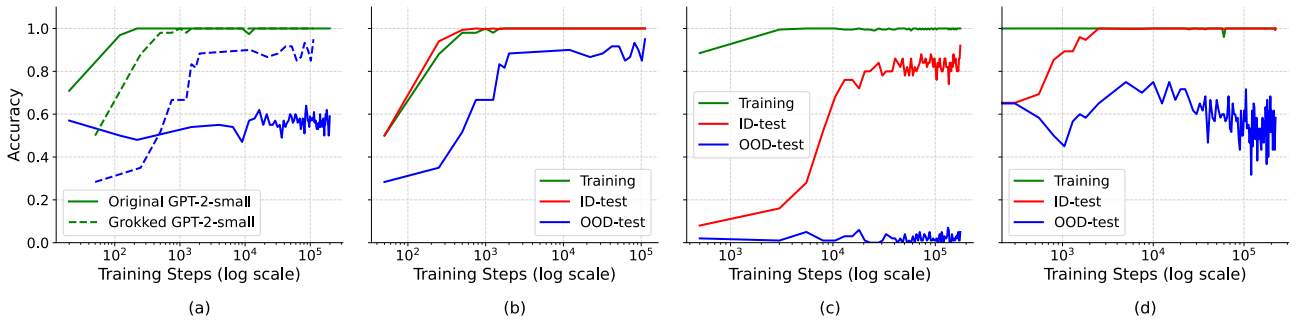

Figure 2. (a) 原始和训练调优后的 GPT2-small 在 OOD 测试集上的比较任务准确率。我们可以观察到，当准确率达到 100 % 时，OOD 准确率几乎保持不变，但之后的差异变得明显。与原始版本不同，经过训练调优的 Transformer 在训练过程中准确率继续提高。(b) 结构化比较任务的训练曲线。IID 和 OOD 行为相似。(c) 结构化组合任务。我们看到近乎完美的 ID 准确率，但在 OOD 测试准确率中没有出现后期跃升。(d) 非结构化（完整段落 Wikipedia）比较设置的训练曲线。尽管 ID 准确率仍显著提高，但复杂性减慢了收敛速度并限制了 OOD 提升。

减少这种开销的技术，例如 Lee et al. (2024) 描述的方法，是至关重要的。具体来说，未来的工作可能会探索：

- 比例定律：确定模型大小、数据集大小和比例 ϕ 如何共同影响训练成本。
- 加速收敛：应用课程学习或专门的优化器缩短“记忆阶段”并加快泛化的开始。
- 预训练：预训练模型可以通过利用先验知识促进理解，提高性能，并加速从记忆到泛化的过渡。由于它们已经编码了基本模式（例如，语言或数学规则），可能需要较少的训练时间来实现泛化。

总而言之，尽管我们证明了在多跳事实问答中使用 grokking 的有效性，但仍有充足的空间来改进、扩展和更好地解释这些新兴能力。我们希望我们的工作能够为未来探索更强大、更透明和更高效的大型语言模型中的隐性推理形式奠定基础。

6 结论

我们已经证明，精心设计的数据合成可以重新塑造事实语言语料库的分布，从而解锁基于悟性的泛化。即使是中等规模的 GPT-2 模型，通过利用后期内部电路的形成，也可以在多跳推理中取得显著的提升——超越那些没有收到合成数据增强的更强大模型。此外，我们的实验证明，事实性的准确性并没有显著降低；平均而言，当模型被提供一个真实与合成事实的良好平衡混合时，它的答案会更加准确。这项工作的主要信息是，通过合成数据提升推断与原子事实的比率 ϕ_r ，仍然是实现新兴推理电路的最直接途径。

然而，我们的方法也强调了一些局限性。要实现所有关系的全面泛化，需要充分增加每个关系的原子事实，这对稀有或低频的关系来说可能具有挑战性。在许多现实世界的语料库中，知识图谱不仅稀疏，而且可能是断开的或部分非注射性的，限制了模型可以从中学习的多跳路径的数量。

最后，自然语言在实际情境中仍然存在挑战。现实世界的文本常常包含含糊的指代、不均匀分布的关系以及不连贯的子图，这使得高质量的数据增强变得不简单。然而，我们的工作展示了隐式推理的潜力——一旦推断事实的比率超过某个临界值，模型会内化能够处理复杂多跳查询的稳健逻辑电路。我们希望这些发现能鼓励在高效合成方法、更广泛领域应用以及对大型语言模型中悟性背后机制的更深入分析上的进一步研究。

7 影响声明

通过展示如何通过有针对性的数据增强来解锁新兴的多跳推理能力，这项工作为更健壮、可解释和高效的知识密集型自然语言处理奠定了基础。

在真实世界事实数据集上诱导“理解”的能力承诺了更广泛的应用，从高风险领域（例如，医疗、法律、教育）到日常问答。同时，这项工作强调了对合成事实进行谨慎策划以防止误导信息的重要性。这种增强推理能力与事实准确性之间的平衡标志着向可信任的、可推广的语言模型迈出的关键一步。

References

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proceedings of the National Academy of Sciences , 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1903070116.

Davies, X., Langosco, L., and Krueger, D. Unifying grokking and double descent. CoRR , abs/2303.06173, 2023. doi: 10.48550/ARXIV.2303.06173. URL https://doi.org/10.48550/arXiv.2303.06173.

Ho, X., Duong Nguyen, A.-K., Sugawara, S., and Aizawa, A. Constructing A Multi-hop QA Dataset

for Comprehensive Evaluation of Reasoning Steps. In Proceedings of the 28th International Conference on Computational Linguistics , pp. 6609–6625, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://www.aclweb.org/anthology/2020.coling-main.580.

Humayun, A. I., Balestriero, R., and Baraniuk, R. G. Training dynamics of deep network linear regions. CoRR , abs/2310.12977, 2023. doi: 10.48550/ARXIV.2310.12977. URL https://doi.org/10.48550/arXiv.2310.12977.

Lee, J., Kang, B. G., Kim, K., and Lee, K. M. Grokfast: Accelerated Grokking by Amplifying Slow Gradients, June 2024. URL http://arxiv.org/abs/2405.20233. arXiv:2405.20233 [cs].

Liu, L., Du, B., Xu, J., Xia, Y., and Tong, H. Joint Knowledge Graph Completion and Question Answering. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining , KDD '22, pp. 1098–1108, New York, NY, USA, August 2022. Association for Computing Machinery. ISBN 978-1-4503-9385-0. doi: 10.1145/3534678.3539289. URL https://dl.acm.org/doi/10.1145/3534678.3539289.

Liu, Z., Michaud, E. J., and Tegmark, M. Omnigrok: Grokking beyond algorithmic data. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023 . OpenReview.net, 2023. URL https://openreview.net/pdf?id=zDiHoIWa0q1.

Loshchilov, I., Hutter, F., et al. Fixing weight decay regularization in adam. arXiv preprint arXiv:1711.05101 , 5, 2017.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020 . OpenReview.net, 2020. URL https://openreview.net/forum?id=B1g5sA4twr.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023 . OpenReview.net, 2023. URL https://openreview.net/pdf?id=9XFSbDPmdW.

Pezeshki, M., Mitra, A., Bengio, Y., and Lajoie, G. Multi-scale feature learning dynamics: Insights for double descent. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA , volume 162 of Proceedings of Machine Learning Research , pp. 17669–17690. PMLR, 2022. URL https://proceedings.mlr.press/v162/pezeshki22a.html.

Plaat, A., Wong, A., Verberne, S., Broekens, J., Stein, N. v., and Back, T. Reasoning with Large Language Models, a Survey, July 2024. URL http://arxiv.org/abs/2407.11511. arXiv:2407.11511 [cs].

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets. CoRR , abs/2201.02177, 2022. URL https://arxiv.org/abs/2201.02177.

Thilak, V., Littwin, E., Zhai, S., Saremi, O., Paiss, R., and Susskind, J. M. The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon. CoRR , abs/2206.04817, 2022. doi: 10.48550/ARXIV.2206.04817. URL https://doi.org/10.48550/arXiv.2206.04817.

Trivedi, H., Balasubramanian, N., Khot, T., and Sabharwal, A. MuSiQue: Multihop Questions via Single-hop Question Composition. Transactions of the Association for Computational Linguistics , 10:539–554, May 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00475. URL https://doi.org/10.1162/tacl_a_00475.

Varma, V., Shah, R., Kenton, Z., Kramár, J., and Kumar, R. Explaining grokking through circuit efficiency, September 2023. URL http://arxiv.org/abs/2309.02390. arXiv:2309.02390 [cs].

Wang, B., Yue, X., Su, Y., and Sun, H. Grokked Transformers are Implicit Reasoners: A Mechanistic Journey to the Edge of Generalization, May 2024. URL http://arxiv.org/abs/2405.15071. arXiv:2405.15071 [cs].

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering, September 2018. URL http://arxiv.org/abs/1809.09600. arXiv:1809.09600 [cs].

A 附录

A.1 引理 1 证明略图

Proof. 我们将论证分为三个部分。

通过选择一个 $(n+1)$ 个不同实体的有序元组可以确定一个简单的 n 跳路径。选择 $v_0, v_1, ..., v_n$ 个不同节点的方式数量为:

这里，$\binom{|\mathcal{V}|}{n+1}$ 是选择 $n+1$ 个不同节点（无序）的方式数量，而 $(n+1)!$ 是将这些节点排序成可能的 n 长度的有向路径的方式数量。

在随机图假设下，从 v_{i-1} 到 v_i 有一个有向边的概率（对于 $i = 1, ..., n$ ）是 $\frac{b}{|\mathcal{V}|-1}$ 。假设边之间独立，所有 n 条边同时存在的概率为

通过期望的线性（应用于每个 $\binom{|\mathcal{V}|}{n+1}(n+1)!$ 有序节点元组），$|\mathcal{F}_n|$ 的期望值为:

对于大的 $|\mathcal{V}|$ ，可以使用二项式近似 $\binom{n}{k} \leq \frac{n^k}{k!}$ 代替 $\binom{|\mathcal{V}|}{n+1}$ ，例如

根据定义 2.3 和 ?? ，

结合第三步的结果为 $\mathbb{E}[|\mathcal{F}_n|] = \binom{|\mathcal{V}|}{n+1}(n+1)!\left(\frac{b}{|\mathcal{V}|-1}\right)^n$ ，我们得到，

这给出了渐近上界，

因此 $\phi_{n,r}$ （也因此 ϕ_n 总体）保持受限于 b^{n-1} ，完成了证明。 □

A.2 引理 2 的证明

Sketch of Proof. 根据定义 2.3 和 ?? ，

$$\phi_{n,r} = \frac{|\mathcal{F}_{n,r}|}{|\mathcal{F}_{A,r}|} \text{ and } |\mathcal{F}_{A,r}| = |\mathcal{V}| \cdot b_r.$$

。由 A.1 ，我们得到

$$|\mathcal{F}_{n,r}| \lesssim \binom{|\mathcal{V}|}{n+1}(n+1)!\left(\frac{b_r}{|\mathcal{V}|-1}\right)^n.$$

，因为 $\frac{b_r}{|\mathcal{V}|-1}$ 是随机选择的一条边属于关系 r 的近似概率。因此，

$$\begin{aligned}\phi_{n,r} &= \frac{|\mathcal{F}_{n,r}|}{|\mathcal{F}_{A,r}|} = \frac{|\mathcal{F}_{n,r}|}{|\mathcal{V}|\, b_r} \\ &\lesssim \binom{|\mathcal{V}|}{n+1}\frac{(n+1)!\, b_r^{n-1}}{|\mathcal{V}|\,(|\mathcal{V}|-1)^n}.\end{aligned}$$

。完整的泛化需要对所有 r 的 $\phi_{n,r} \geq \phi_G$ ，从而得出所需的约束

$$\begin{aligned}&\phi_G \leq \phi_{n,r} \\ \Leftrightarrow\ &\phi_G \leq \binom{|\mathcal{V}|}{n+1}\frac{(n+1)!\, b_r^{n-1}}{|\mathcal{V}|\,(|\mathcal{V}|-1)^n} \\ \Leftrightarrow\ &b_r \geq \sqrt[n-1]{\frac{\phi_G\,|\mathcal{V}|\,(|\mathcal{V}|-1)^n}{\binom{|\mathcal{V}|}{n+1}\,(n+1)!}}.\end{aligned}$$

。换句话说，如果任何关系 r 的 b_r （其分支因子）未能超过此阈值，那么 $\phi_{n,r}$ 无法达到 ϕ_G 。因此，该特定关系将不会“掌握”，因此 KB 无法完全概括于 n 跳事实（尽管对于其他关系，它仍可能部分地概括）。 □

A.3 节点计数界的形式推导

类似于 A.2 ，对于完全泛化，我们可以为节点计数 $|\mathcal{V}|$ 推导出一个界限。根据要求 $\forall r \in \mathcal{R}, \phi_{n,r} \geq \phi_G$ ，我们得到，

$$\forall r \in \mathcal{R}, \phi_{n,r} \geq \phi_G$$
$$\Leftrightarrow \min_r \binom{|\mathcal{V}|}{n+1} \frac{(n+1)! b_r^{n-1}}{|\mathcal{V}|(|\mathcal{V}|-1)^n} \geq \phi_G$$

换句话说，最差的可泛化关系 r 仍然需要满足 $\phi_{n,r} \geq \phi_G$ ，以便 KB 能够完全泛化。在不使用二项式近似的情况下，解决 $|\mathcal{V}|$ 后得到，

$$\begin{aligned}
&\min_r \binom{|\mathcal{V}|}{n+1} \frac{(n+1)! b_r^{n-1}}{|\mathcal{V}|(|\mathcal{V}|-1)^n} \geq \phi_G \\
\Leftrightarrow &\min_r \frac{|\mathcal{V}|!}{(n+1)!(|\mathcal{V}|-n-1)!} \frac{(n+1)! b_r^{n-1}}{|\mathcal{V}|(|\mathcal{V}|-1)^n} \geq \phi_G \\
\Leftrightarrow &\min_r \frac{|\mathcal{V}|!}{(|\mathcal{V}|-n-1)!} \frac{b_r^{n-1}}{|\mathcal{V}|(|\mathcal{V}|-1)^n} \geq \phi_G \\
\Leftrightarrow &\frac{(|\mathcal{V}|-1)!}{(|\mathcal{V}|-n-1)!(|\mathcal{V}|-1)^n} \min_r b_r^{n-1} \geq \phi_G \\
\Leftrightarrow &\frac{(|\mathcal{V}|-1)!}{(|\mathcal{V}|-n-1)!(|\mathcal{V}|-1)^n} \geq \max_r \frac{\phi_G}{b_r^{n-1}} \\
\Leftrightarrow &\frac{\Gamma(|\mathcal{V}|)}{\Gamma(|\mathcal{V}|-n)(|\mathcal{V}|-1)^n} \geq \max_r \frac{\phi_G}{b_r^{n-1}}
\end{aligned}$$

因此我们有：

$$|\mathcal{V}| \geq \min\left\{v \in \mathbb{N} : \frac{\Gamma(v)}{\Gamma(v-n)\,(v-1)^n} \geq \max_r \frac{\phi_G}{b_r^{n-1}}\right\}.$$

对于一个实证例子（即，使用随机生成的图），参见图 3 。

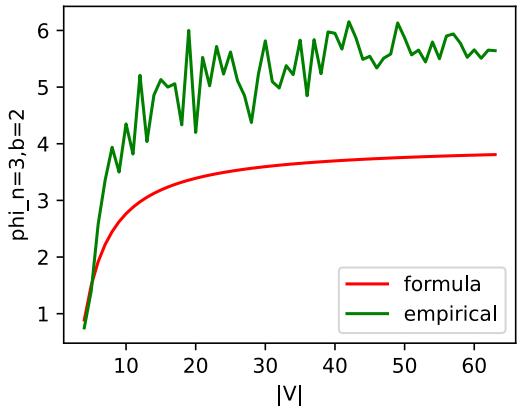

Figure 3. $b_r = 2$ 的 $\phi_{3,r}$ (y 轴) 随 $|\mathcal{V}|$ (x 轴) 的增长。红线是 $\phi_{3,r} = \binom{|\mathcal{V}|}{4} \frac{96}{|\mathcal{V}|(|\mathcal{V}|-1)^3}$ (公式值)。绿线是通过随机生成具有相同数量的节点 ($|\mathcal{V}|$) 和边 ($|\mathcal{V}|b_r$) 的图得到的经验值。由于随机性，我们生成的/经验图可以在局部具有更高的分支因子，从而导致整体上推断出更多事实。因此，我们的 $\phi_{3,r}$ 公式实际上低估了 $\phi_{3,r}$ 的真实值。尽管如此，该公式正确地近似了形状和数量级。这对于其他 b_r 和 n 的组合也是成立的。

A.4 定性示例

在以下部分中，我们展示了模型未能正确分类的组成和比较任务中 QA 对（“问题”，“标准答案”）的定性例子。一些错误源于 2WikiMultiHopQA 数据集中的不一致性，另外一些则来自我们的增强策略。总体而言，我们相信模型具有实现 100 % 精度的能力，正如之前的研究所证明。组成 2WikiMultiHopQA 数据集中存在一些语法上的不一致，使我们的模型无法达到 100 % 精度。

国籍问题可能会有不一致的真实格式 对于有关国籍的问题，标准答案可以是一个形容词或国家的名称，即使后者在语法上不正确。

问题：威廉·西摩，第三代萨默塞特公爵的父亲是什么国籍?
真实答案：英国人

问题：萨伏依公爵阿马迪乌斯八世的母亲是哪国人?
标准答案：法国

在国家相关的回答中用形容词代替名词 在某些情况下，标准答案是形容词，而不是指代国家的名词。

问题：Paper Bird 的小号手来自哪个国家?
标准答案：美国

比较比较错误主要是由于数据增强算法的限制，该算法未能为某些关系查询生成足够数量的推断事实。大多数错误与湖泊、河流或机场有关，我们认为这是由于基础数据的稀疏性所致。

问题：Wainwright/Wainwright Field 21 机场和 Roberval Air Saguenay 水上机场是否位于同一国家?
事实依据：是的

问题：安大略省东费里斯的长湖和萨斯喀彻温省的蒙特利尔湖是否都位于同一个国家?
基本事实：是的

问题：鹿溪、奥塞奇河和大草原狗溪是否位于同一个国家?
答案：是的

问题：Chicoutimi/Saint-Honoré 机场和 Rtishchevo 空军基地是否位于同一个国家?
真实情况：不是

A.5 数据综合

在这里，我们展示了在整个数据增强过程中使用的提示。在实验中，我们利用了 GPT-4o 和 o1-mini 来生成多样且高质量的增强数据。

A.5.1 成分

图解析

> You are graph gpt. You build graph based on the provided text. Find all objects, their relations and types.
> Pick one of the following types: - Person - Location - Object (include everything that was not above)
> Return the following format with numbering: 1. <Avatar; Film><director><James Cameron; Person> 2. <James Cameron; Person><directed><Titanic; Object>

问题格式化

> You are a question formatting assistant. Your task is to create questions based on the given relations and objects.
> Use the provided examples as a guide for the question style. Ensure that the answer remains unchanged and enclosed in <a> tags. You may rephrase one question, given the example format. Strictly follow the logic of given examples. Connect it in the following logic: <obj1> -> <rel1> -> <rel2> -> <obj3>
> Return numbered responses in format: 1. What is the director of the film that James Cameron produced?<a>Steven Spielberg 2. Who directed the movie starring Tom Cruise?<a>Christopher Nolan

A.5.2 比较

原子事实生成

> You are a helpful assistant that generates geographical facts. Generate new unique locations and their countries in the following format: Follow the style of the examples, but do not use the same locations.
> Rules: 1. Use real locations and countries 2. Each location should be unique 3. DO NOT REUSE PROVIDED EXAMPLES 4. Do not answer the question - only provide locations 5. Do not use formatting except for numbering 6. Generate equal amount of NEW!!! locations for following countries:

详细原子事实生成

You are a helpful assistant that generates geographical facts. Based on the provided examples, generate a paragraph for each location-country pair. Strictly follow the style and lenght of the provided examples Do not answer the question - only provide the paragraph with numbering. DO not return empty lines. One by one. Return the number according to the given data. Here are the examples: