arXiv:2601.00942v1 [cs.LG] 2 Jan 2026

随机性下的可靠性：不同解码温度下稀疏与稠密语言模型行为的实证分析

Kabir Grover

2025年12月

摘要

稀疏混合专家（MoE）架构在大型语言模型中的日益普及，引发了关于其在随机解码下可靠性的重要问题。尽管条件计算能够显著提升计算效率，但稀疏路由与基于温度的采样之间的相互作用是否会损害输出稳定性，使其逊于稠密架构，目前仍不清楚。本研究探讨 MoE 模型中的条件计算是否会放大解码诱发的随机性，进而导致可靠性随着温度升高而降低。我们在答案可客观验证的确定性算术推理任务上，评估了三种具有代表性的模型：OLMoE-7B（稀疏基础模型）、Mixtral-8x7B（稀疏指令微调模型）和 Qwen2.5-3B（稠密指令微调模型）。实验涵盖四种解码配置，从贪心解码到 T=1.0。评估指标包括准确率、格式合规性、重复生成结果的一致性以及置信度指标，共计 9,360 次模型生成。结果表明，在所有解码温度下，稀疏指令微调模型的稳定性都与稠密指令微调模型相当，而稀疏基础模型则随着温度升高而呈现系统性退化。这些发现表明，在确定性任务中，决定模型对解码随机性鲁棒性的主要因素是指令微调，而非架构稀疏性。我们讨论了这些结果对于在可靠性关键型应用中部署稀疏语言模型的启示，并重点指出了可在不牺牲输出稳定性的情况下安全采用稀疏架构的情形。

目录

1 引言 3
1.1 研究目标 .3
1.2 主要发现 .3

2 背景 3
2.1 自回归语言模型生成 .3
2.2 温度缩放 .4
2.3 混合专家架构 .4
2.4 计算效率：稠密与稀疏架构 .5
2.5 指令微调 .5

3 实验设置 5
3.1 模型.5
3.2 任务 .6
3.3 解码配置 .6
3.4 评估指标 .6

4 结果 7
4.1 总体准确率 .7
4.2 架构比较：稀疏模型与稠密模型 .8
4.3 格式合规性.8
4.4 失败模式分析 .9
4.5 特定任务分析：变量绑定 .10
4.6 特定任务分析：多步算术 .11
4.7 一致性分析 .11
4.8 置信度分析 .12
4.9 结果总结 .12

5 讨论　讨论 1 3
5.1 初始假设评估 .13
5.2 实验混杂因素 .14
5.3 OLMoE 退化的机制解释.14
5.4 概率集中假设 .14
5.5 局限性 .15
5.6 稳健发现 .15

6 结论 15
6.1 主要结论.16
6.2 实践建议 .16
6.3 未来工作 .16

1 引言

在生产环境中部署语言模型，必须仔细权衡确定性解码策略与随机解码策略。贪心解码可确保结果可复现，而温度采样则能增加多样性，这可能有益于某些应用。然而，对于需要精确输出的任务，包括数学推理、结构化生成和事实检索，采样引入的随机性可能会损害可靠性。

随着**稀疏混合专家（MoE）模型**的出现，这一考量显得尤为重要。与每个词元都会激活所有参数的稠密 Transformer 不同，MoE 架构会将每个词元路由至一组专门的“专家”子网络。例如，Mixtral-8x7B 的总参数量为 467 亿，但每个词元仅激活约 129 亿个参数，因此能够以大幅降低的计算成本实现具有竞争力的性能。

稀疏路由机制引出了一个根本性问题：*MoE* 模型中的条件计算是否会与温度采样产生不利交互？如果路由决策对微小扰动很敏感，而温度又会给词元选择引入随机性，这些效应可能相互叠加，使稀疏模型的可靠性低于对应的稠密模型。

1.1 研究目标

本研究将探讨以下假设：

> **随着解码温度升高，稀疏 MoE 模型的性能退化是否会加速，或者其行为是否会与稠密模型呈现出定性差异？**

为回答这一问题，我们设计了受控实验，采用确定性算术任务，具体而言，是答案明确无歧义的问题，以便精确测量多个实验运行中的准确率、格式合规性和一致性。

1.2 主要发现

本研究的核心发现可概括如下：**温度敏感性主要由指令微调决定，而非架构稀疏性**。

稀疏指令微调模型（Mixtral）表现出与稠密指令微调模型（Qwen）相当的稳定性，两者从贪心解码到 $T = 1.0$ 都保持了准确率一致。唯一表现出随温度变化而退化的模型是稀疏基础模型 （OLMoE），它未经过指令微调。

这些结果表明，指令微调能够赋予模型抵御温度变化的鲁棒性，且这一效果不受架构影响。对于经过良好调优的模型，在确定性任务上，MoE 模型固有的路由方差似乎不会与采样随机性叠加。

本文其余部分安排如下：第 2 节介绍语言模型解码、温度缩放和 MoE 架构的技术背景。第3节详述实验方法。第 4 节呈现实证结果。第5 节讨论其影响和局限性。第6 节总结并为从业者提出建议。

2 背景

2.1 自回归语言模型生成

自回归语言模型按顺序生成文本，每次生成一个词元。给定一个词元序列 x_1，x_2，……，x_t，模型会预测词汇表中各词元的概率分布，

后续词元：

$$P(x_{t+1} \mid x_1, \ldots, x_t) \tag{1}$$

This 该分布通过对模型输出的 logit 应用 softmax 函数得到 s:

$$P(x_{t+1} = v) = \frac{\exp(z_v)}{\sum_{v'} \exp(z_{v'})} \tag{2}$$

其中 z_v 表示词汇项 v的 logit（softmax 前的分数）。

2.2 温度缩放

温度缩放会在采样前调整输出分布。设温度参数为 T，则概率变为：

$$P_T(x_{t+1} = v) = \frac{\exp(z_v/T)}{\sum_{v'} \exp(z_{v'}/T)} \tag{3}$$

温度参数会产生以下效果：

- $T \to 0$ **（贪心解码）：** 分布会变得越来越尖锐，概率最高的词元几乎获得全部概率质量，从而实现确定性选择。

- $T = 1$**：** 模型原始学习到的分布未经修改，直接用于生成。
- $T > 1$**：** 分布变得更加平坦，使排名较低的词元被选中的概率上升。

Holtzman 等人 [2] 证明，从完整分布中进行采样（即使在 $T = 1$下）也可能因不可靠的尾部而生成退化文本；所谓不可靠的尾部，是指概率极低却偶尔会被选中的词元。top-k 和核采样（top-p）等方法通过截断分布来解决这一问题。温度仍是控制多样性与确定性之间权衡的主要机制。

本研究评估了四种温度配置：贪心（$T = 0$）、 T =0.1、 $T = 0.5$和 $T = 1.0$，涵盖从完全确定性解码到使用模型未经修改的分布。

2.3 混合专家架构

混合专家（MoE）架构 [5, 1] 将 Transformer 块中的稠密前馈网络（FFN）替换为多个并行的“专家”网络和一种路由机制。

标准 Transformer 前馈网络（FFN）的计算方式为：

$$\mathrm{FFN}(x) = W_2 \cdot \mathrm{ReLU}(W_1 \cdot x) \tag{4}$$

包含 N 个专家的 MoE 层则计算：

$$\mathrm{MoE}(x) = \sum_{i=1}^{N} G(x)_i \cdot E_i(x) \tag{5}$$

其中， E_i 表示第 i 个专家网络， $G(x)$ 是一个门控函数，用于生成稀疏权重向量；通常只有排名前 k 的专家会获得非零权重。

核心见解是，模型在训练期间可以访问所有专家，但在推理期间只会使用其中一部分。这使得模型容量（总参数量）得以扩展，而无需按比例增加计算量（激活参数）。

Mixtral-8x7B 每个 MoE 层采用 8 个专家和 top-2 路由，每个词元会激活 8 个专家中的 2 个。总参数量：47B。每个词元的激活参数：约 13B。

2.4 计算效率：稠密与稀疏架构

从推理计算的角度来看：

- **稠密模型：** 推理浮点运算次数 $\approx 2 \times$ 每个词元的参数量
- **稀疏混合专家模型：** 推理浮点运算次数 $\approx 2 \times$ 激活参数 $= 2 \times$总参数量$\times$ $(k \ / \ N)$

其中，k 表示激活专家的数量，N 表示专家总数。

这种计算效率使 MoE 架构备受关注。然而，评估效率时也必须考虑可靠性。一个潜在的问题是：稀疏路由是否会引入与采样随机性叠加的方差？如果隐藏状态的微小扰动导致专家选择发生变化，那么温度采样与专家路由相结合，可能会产生一致性较低的输出。

2.5 指令微调

指令微调（或对齐）训练模型遵循用户指令，而不仅仅是预测下一个词元。基础模型学习文本补全；指令微调模型学习执行指定任务。

这一区别与当前的实验设计相关，因为模型被要求“仅输出最终数字”。基础模型可能会忽略这一指令，并提供解释性推理。指令微调模型则应遵循指定格式。先前研究表明，即使是能力较强的模型，在格式合规性方面也面临挑战：Zhou 等人 [6] 发现，GPT-4 在格式合规性基准测试中的达标率仅约为 80%，而较小的模型通常只有 30–40%。

其　基础模型与指令微调模型之间的区别是我们研究发现的核心。

3 实验设置

3.1 模型

我们评估 涵盖稀疏模型和稠密架构的三种基于Transformer的语言模型 s:

模型	架构	参数总量	激活参数	Type
OLMoE-7B	稀疏混合专家模型	6.9B	$\approx$ 1.3B	Base
Qwen2.5-3B	稠密	3.1B	3.1B	指令微调
Mixtral-8x7B	稀疏混合专家模型（8 个专家，top-2）	47B	$\approx$ 13B	指令微调

表 1：本研究评估的模型。

OLMoE-7B [4]是一种稀疏混合专家模型，拥有 69 亿个参数，每个词元激活的参数约为 10 亿。研究采用其基础模型 版本（未经指令微调），以考察未对齐情况下不同温度下的稀疏模型行为。

Qwen2.5-3B 是一种拥有 30 亿个参数的稠密 Transformer，推理期间所有参数都会被激活。此处采用其指令微调变体，这是一种对齐良好的稠密模型。

Mixtral-8x7B [3] 是一种大型稀疏混合专家模型，每层包含 8 个专家，并采用 top-2 路由。它的总参数量为 467 亿，但每个词元仅激活约 129 亿个参数，因此能够以更低的计算成本实现出色的性能。此处采用其指令微调变体。

由于显存限制，OLMoE 和 Mixtral 使用 4 位量化（bitsandbytes NF4）运行。Qwen2.5-3B 则以全精度运行。

计算资源：实验在配备 40GB 显存的 NVIDIA A100 GPU 上进行。生成 9,360 个结果的总推理时间约为 3.5 小时。

3.2 任务

本研究采用两项答案可客观验证的确定性推理任务：

变量绑定（120 个示例）。跟踪变量在一系列操作中的赋值过程：

a=3，b=2，c=a+b，d=c+1。d = ？（答案：6）

多步算术（120 个示例）。计算需要遵循运算顺序的表达式：

```
(45-12)*3+7 = ?    (Answer: 106)
```

这两项任务都有明确无歧义的正确答案。提示指示模型<only>仅</only>输出最终数字，以便测量准确率和格式合规性。

3.3 解码配置

设置	采样	温度	Runs
贪心	No	–	1
T=0.1	Yes	0.1	4
T=0.5	Yes	0.5	4
T=1.0	Yes	1.0	4

表 2：评估的解码配置。

贪心解码是确定性的，只需运行一次。采样配置使用多次运行来评估一致性。不应用 top-k 或 top-p 截断；温度是唯一的随机性来源。

实验总规模：3 个模型 $\times$ 240 个示例 $\times$ 13 次运行 $= 9$，360 个生成结果。

3.4 评估指标

准确率。从每个输出中提取第一个整数，并与真实值进行比较。评估标准较宽松：“答案是 106”会被判为正确（尽管不符合合规性要求）。

合规性。对输出是否仅由数值构成进行二元评估。“106”符合要求；“106。以下是我的推理……”则不符合要求。

一致性。在每个温度设置下，准确率在各次运行中的标准差。数值越低，表示行为越一致。

置信度。生成词元的平均对数概率。数值越负，表示模型置信度越低。

4 结果

4.1 总体准确率

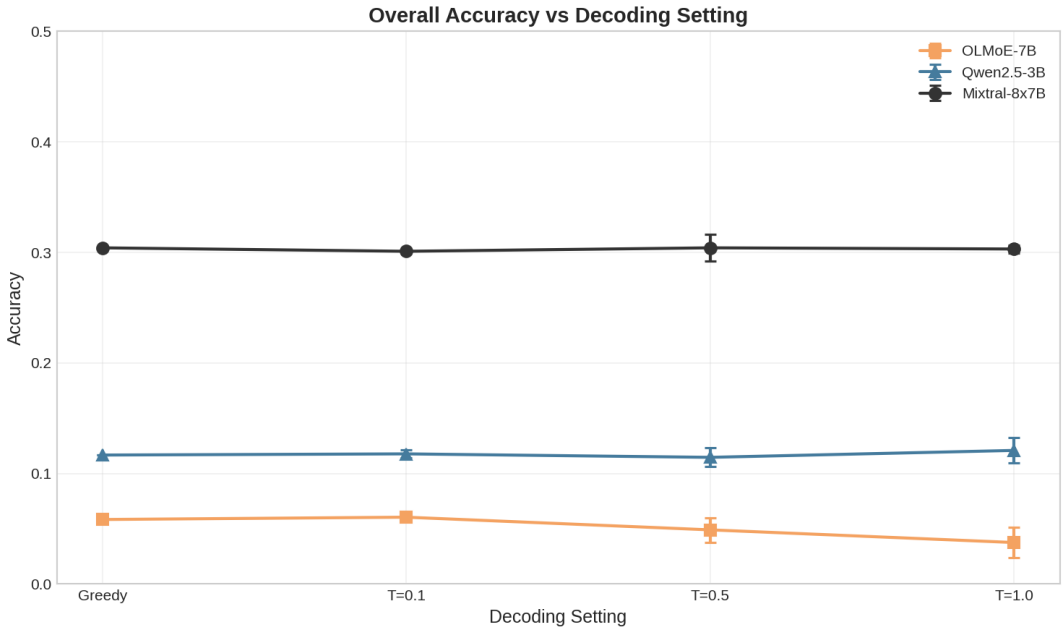


图1：不同解码配置下的总体准确率。误差线表示各次运行的标准差。

图 1展示了所有三个模型的准确率随温度变化的情况。横轴表示从贪心（确定性）到 $T = 1.0$ （最大随机性）的温度设置。误差线表示重复运行间的方差。

模型	贪心	T=0.1	T=0.5	T=1.0	趋势
OLMoE-7B	5.8%	6.0%	4.9%	3.8%	↓ 退化
Qwen2.5-3B	11.7%	11.8%	11.5%	12.1%	→ 稳定
Mixtral-8x7B	30.4%	30.1%	30.4%	30.3%	→ 稳定

表 3：各模型在不同温度配置下的准确率。

OLMoE-7B 呈现出明显的下降趋势，准确率从贪心解码时的 5.8% 降至 3.8

Qwen2.5-3B 性能稳定，在所有配置下准确率都仅在 11–12% 左右小幅波动。温度对该模型的性能影响可以忽略不计。

Mixtral-8x7B 同样保持稳定的性能，准确率始终维持在约 30%。尽管它是规模最大的模型，且路由机制最为复杂，但并未出现退化。

值得注意的是，尽管 Mixtral 采用稀疏架构，其准确率仍然最高。这一观察提供了初步证据，表明稀疏性本身并不会损害可靠性；如果架构稀疏性存在问题，那么预计 Mixtral 会表现出最低的稳定性，而非最高的准确率。

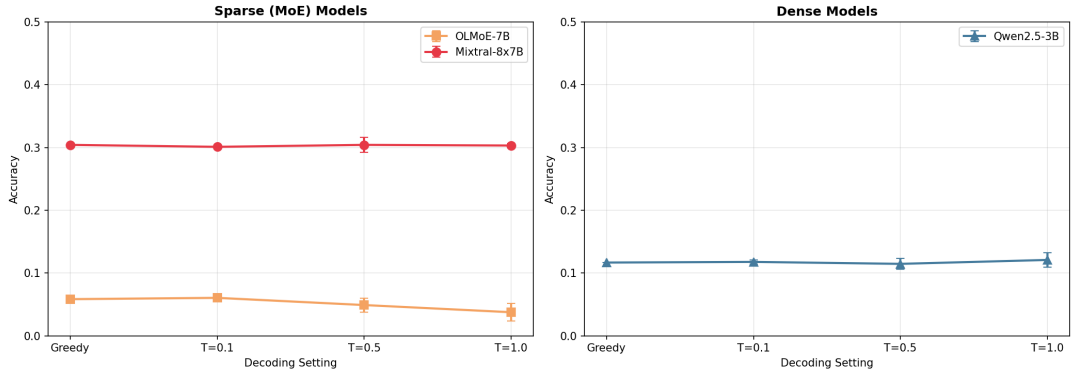


图 2：稀疏模型（左图）与稠密模型（右图）的比较。OLMoE 出现退化，而 Mixtral 保持稳定。

4.2 架构比较：稀疏模型与稠密模型

图 2 展示了按架构类型划分的准确率数据。左侧面板显示两个稀疏混合专家模型（OLMoE 和 Mixtral）；右侧面板显示稠密模型（Qwen）。

如果稀疏架构会导致温度敏感性，那么预期左图中的两个模型都会表现出性能下降。但实证结果与这一预期相矛盾：

- **左图（稀疏模型）：** Mixtral 在所有温度下都保持稳定；OLMoE 则随着温度升高而出现退化。
- **右图（稠密模型）：** Qwen 在所有温度下都保持稳定的性能。

关键观察在于，这两个稀疏模型的表现并不一致。如果稀疏性本身会导致温度敏感性，那么预期 OLMoE 和 Mixtral 都会出现退化。OLMoE（出现退化）与其他模型（保持稳定）之间的区别并非架构所致：OLMoE 是一个缺少指令微调的**基础模型**，而 Mixtral 和 Qwen 都经过了**指令微调**。

4.3 格式合规性

模型	贪心	T=0.1	T=0.5	T=1.0
OLMoE-7B	0%	0%	0%	∼1%
Qwen2.5-3B	100%	100%	100%	100%
Mixtral-8x7B	58.8%	58.4%	58.9%	56.7%

表4：合规率（按指令仅输出数值）。

格式合规性结果揭示了三种截然不同的行为模式：

OLMoE-7B 的**合规率为 0%**。在 3,000 多条生成结果中，该模型从未生成过仅包含数值的输出。每条回复都包含解释、推理步骤或补充文本。作为未经指令微调的基础模型，它无法理解并执行格式规范。

Qwen2.5-3B 达到**100% 的合规性**。每条回复都只包含一个数值。即使回答错误，模型仍遵守了指定格式。这表明模型完全理解并执行了指令。

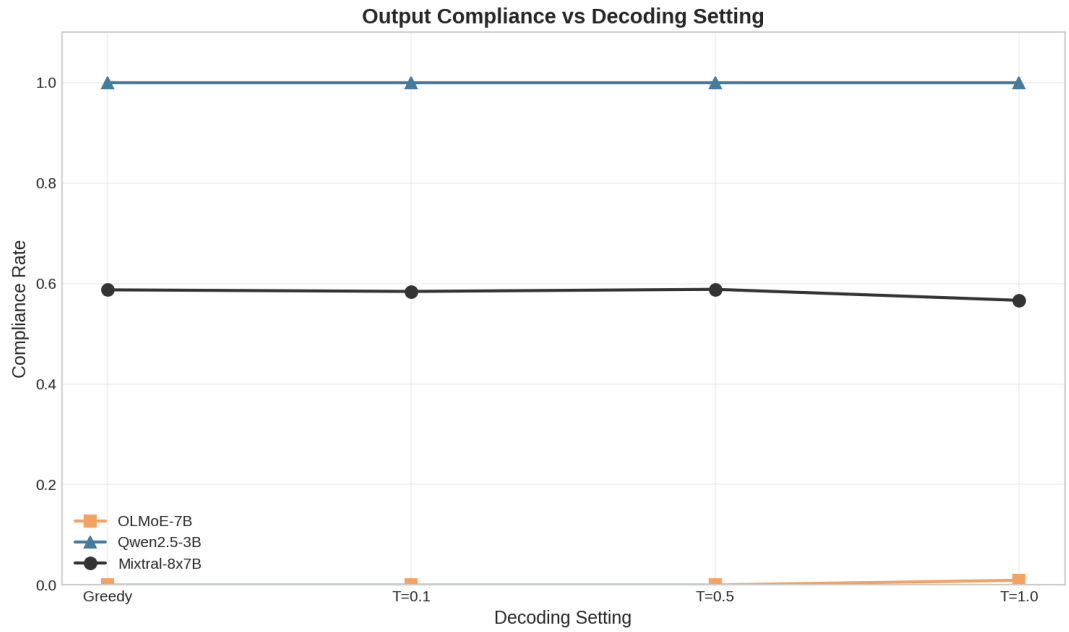


图3：不同解码配置下的输出格式遵循度。

Mixtral-8x7B 的合规性**约为 58%**。该指令微调模型通常会遵守格式，但约 42% 的回复中会包含补充文本。在 $T = 1.0$ （56.7%）时，合规性略有下降，这表明随机性略微增加会提高格式违规的概率。

格式合规性有助于解释观察到的准确率模式。OLMoE 的合规率为 0%，导致输出冗长，出错的机会也随之增加。Qwen 的合规率为 100%，因此输出简洁且聚焦。

4.4 失败模式分析



图4：不同模型和解码配置下的失败类型分布。

图4 展示了失败模式分类体系，分类如下：

- **数值错误：**格式正确，但数值有误
- **额外文本：**数值答案之外的补充说明或文本

- **幻觉步骤：**捏造的推理或中间计算
- **拒绝回答：** 明确拒绝提供答案
- **不完整：** 输出被截断或为空

OLMoE-7B：失效主要表现为“额外文本”（贪心情况下约占 100%）。随着温度升高，少量幻觉步骤和拒绝回答开始出现，但过度冗长仍是主要失效模式。

Qwen2.5-3B： 100% 的失效被归类为“数值错误”。模型始终遵守格式要求；失效完全是由计算错误导致的。这体现了可预测且特征明确的失效行为。

Mixtral-8x7B：观察到多种失效模式：约 57% 为数值错误，12–17% 为额外文本，约 22% 为幻觉步骤。在不同温度下，失效分布保持相对稳定。

该分析揭示了失效特征之间的一个根本区别：

- **基础模型（OLMoE）：**失败源于指令不遵循，这是一种行为缺陷。
- **指令微调模型（Qwen、Mixtral）：**失败源于计算错误，这是一种能力局限。

4.5 任务特定分析：变量绑定

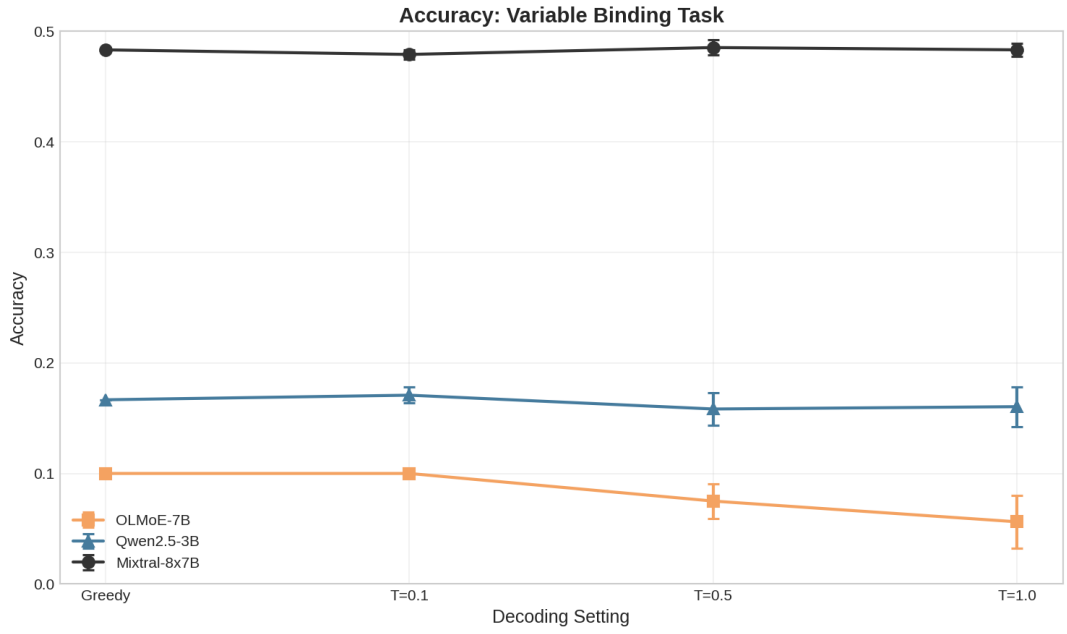


图 5：不同解码配置下变量绑定任务的准确率。

变量绑定要求在多次顺序赋值过程中跟踪状态，这代表着一种不同于纯粹算术计算的认知技能。结果延续了总体分析中观察到的模式：

- **Mixtral-8x7B** 准确率最高，表明稀疏路由能够有效地进行顺序变量跟踪。
- **Qwen2.5-3B** 在不同温度下表现稳定。

- **OLMoE-7B**仍会随着温度升高而出现退化，这表明基础模型的局限性在不同任务类型中依然存在。

4.6 任务特定分析：多步算术

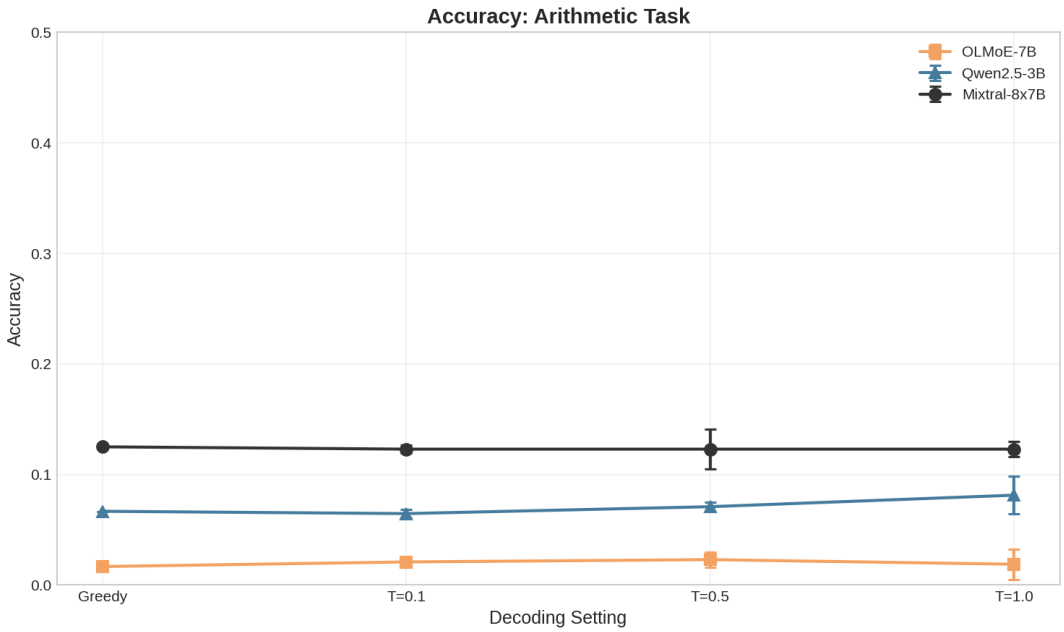


图 6：不同解码配置下多步算术任务的准确率。

多步算术要求保持中间结果，并按照正确的优先级顺序执行运算。结果与变量绑定分析相呼应：

- 整体趋势在不同任务中均一致：指令微调模型保持稳定，而基础模型则出现退化。

- **Mixtral-8x7B**再次取得最高准确率，这表明其较大的活跃参数数量（13B）带来了更强的算术能力。
- **OLMoE-7B**的退化在两种任务类型中表现一致，证实温度敏感性是该基础模型的一般属性，而非特定于某项任务。

4.7 一致性分析

一致性用于量化在相同温度设置下重复运行时准确率的方差。数值越高，表示行为越可预测。

- **贪心解码** ($T = 0$) 对所有模型都能实现完美一致性，这与其确定性相符。

- **指令微调模型**（Mixtral 和 Qwen）即使在 T =1.0 的温度下也能保持较高的一致性，表明训练提升了模型对采样随机性的鲁棒性。
- **OLMoE-7B**的一致性会随温度升高而降低，表明其输出变得越来越难以预测。

在较高温度下，准确率和一致性均有所下降，这使基础模型尤其不适合生产部署。



图 7：不同温度配置下多次运行的一致性（标准差的倒数）。

4.8 置信度分析

置信度以生成词元的平均对数概率来量化，数值越接近零表示置信度越高。

- **温度升高会自然降低置信度**，因为较高的温度会使概率分布变平坦。
- **Mixtral-8x7B** 整体置信度最高，这与其准确率最高相一致。
- **Qwen2.5-3B** 在不同温度下均保持适中且稳定的置信度。
- **OLMoE-7B**的置信度较低，尤其是在温度较高时。

置信度与指令微调质量相关：格式遵循表现更好的模型，其输出置信度也更高。这表明，经过良好对齐的模型具有校准得更好的不确定性估计。

这一观察具有实际意义：词元概率可作为可靠性信号。对于指令微调模型，高置信度输出（平均对数概率 > -0.5）更有可能是正确的。此类基于置信度的筛选对基础模型的效果会较差，因为其置信度与正确性的相关性没有那么强。

4.9 结果总结

- **准确率：**OLMoE 的性能随温度升高而退化（5.8% $\rightarrow$ 3.8%）。Qwen 和 Mixtral 则保持稳定。
- **特定任务性能：**变量绑定和算术均呈现一致的模式：指令微调模型能够抵御温度升高带来的性能退化。
- **架构：** 稀疏与稠密架构并不能决定稳定性。Mixtral（稀疏模型）是评估中最稳定的模型。

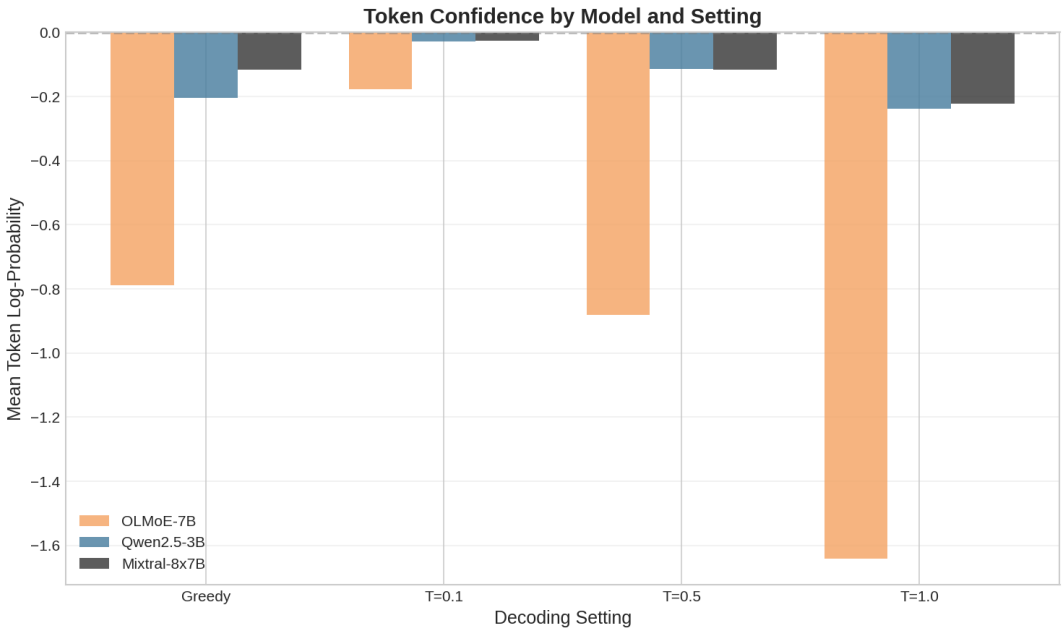


图 8：不同解码配置下的模型置信度（生成词元的平均对数概率）。

- **合规性：** Qwen $= 100\%$，Mixtral $\approx 58\%$，OLMoE $= 0\%$。指令微调是区分因素。
- **一致性：** 指令微调模型在多次运行中的准确率保持一致；OL-MoE 在较高温度下方差增大。
- **置信度：** 较高的置信度与指令微调相关。Mixtral 的置信度最高；OLMoE 的置信度最低。
- **失效模式：** 基础模型因格式不合规而失效。指令微调模型因计算错误而失效。

5 讨论

5.1 初始假设的评估

初始假设提出：**随着解码温度升高，稀疏混合专家模型是否比稠密模型退化得更快?**

根据实证结果，答案是：**温度敏感性取决于指令微调，而非架构。**

观察到的模式如下：

- Mixtral-8x7B（稀疏模型，指令微调）：在所有温度下均保持稳定
- Qwen2.5-3B（稠密模型，指令微调）：在所有温度下均保持稳定
- OLMoE-7B（稀疏模型，基础模型）：随着温度升高而退化

稀疏指令微调模型（Mixtral）的可靠性与稠密指令微调模型（Qwen）相当。唯一表现出退化的是基础模型，而它恰好采用了稀疏架构。

这些结果并不支持“稀疏模型退化得更快”这一说法。相反，在接受评估的模型中，指令微调似乎能够增强模型应对温度变化的鲁棒性，且不受架构选择的影响。

5.2 实验混杂因素

必须承认一项方法论局限：实验设计未纳入稠密基础模型。

	Base	指令微调
稀疏模型	OLMoE（退化）	Mixtral（稳定）
稠密	未评估	Qwen（稳定）

表 5：实验覆盖范围。省略的单元格限制了因果推断。

由于缺少稠密基础模型，无法确定 OLMoE 的退化是否源于：

1. 稀疏架构（架构效应），或
2. 缺乏指令微调（对齐效应），或
3. 两种因素的交互作用

Mixtral 的稳定性（稀疏模型、经过指令微调）表明，稀疏性本身并不会造成问题。然而，要明确排除架构是影响因素之一，还需要一个稠密基础模型。

5.3 OLMoE 退化的机制解释

遵循度分析为解释 OLMoE 的退化模式提供了依据。

OLMoE 的遵循度为 0%；它从未生成过仅包含数值的输出。每个回答都包含解释性内容。该模型不会将“仅输出最终数字”理解为需要执行的指令。

随着温度升高：

- 输出变得更长，变化也更大
- 冗长程度增加，为错误提供了更多机会
- 正确答案被淹没在解释性文本中
- 偶尔会出现拒绝回答或离题内容

指令微调模型能够避免这种失效模式。Qwen 会严格按照要求的格式输出（合规性达 100%）。Mixtral 大多也能做到（约 58%）。简洁且聚焦的输出限制了温度引起的偏移空间。

5.4 概率集中假设

基础模型与指令微调模型之间的温度敏感性差异，或可通过**概率集中**来解释。

当模型对正确响应具有较高置信度时，它会为正确的词元序列分配极高的概率（例如，99%）。即使在 $T = 1.0$时，采样也会频繁选出正确答案。

对于 OLMoE，问题不在于概率集中在错误答案上，而在于模型不理解任务格式。其概率质量分布在多个合理的后续内容上（解释、推理步骤、偏题内容）。温度升高会加剧这种多样性，使输出更难预测。

相比之下，Mixtral 和 Qwen 经过训练，会将概率质量集中在指定行为上：输出一个数值并终止。这种集中的分布表现出对采样噪声的鲁棒性。

5.5 局限性

缺少对比。 由于缺少稠密基础模型，无法将架构效应与指令微调效应完全区分开来。理想的实验设计应纳入 Llama2-7B-base 这样的模型，以补全 2×2析因设计。

模型规模差异。 比较涉及 1B 激活参数（OLMoE）、3B（Qwen）和 13B 激活参数（Mixtral）。规模差异可能部分解释准确率的变化。

任务范围有限。 评估仅涉及算术任务。开放式生成或创意任务可能呈现不同的模式。

单一基础模型。 OLMoE 仅代表一种基础模型。其他基础模型可能表现出不同的行为。

量化效应。 OLMoE 和 Mixtral 使用了 4 位量化，这可能会影响结果。量化可能影响绝对准确率，但模型内部的相对温度趋势应当会保持不变。

5.6 稳健发现

尽管存在方法上的局限性，以下几项发现仍有充分依据：

1.**稀疏指令微调模型可以达到与稠密指令微调模型相当的稳定性**。 Mixtral 未表现出随温度升高而退化的现象；它是评估中最稳定、准确率最高的模型。2. **指令微调会显著影响合规性**。 合规性从 0%（OLMoE）到 100%（Qwen）的差距，体现了对齐对格式遵循的影响之大。3. **不同类型模型的失效模式存在系统性差异**。 基础模型会因不遵循指令而失败（输出过多文本）。指令微调模型则会因计算错误而失败（数值错误）。

对于从业者：部署经过良好调优的模型时，适中的温度（最高至 $T = 1.0$）不太可能显著损害其在确定性任务上的可靠性。模型训练质量比架构选择更为重要。

6 结论

本研究探讨了这样一个问题：随着解码温度升高，稀疏混合专家模型是否会比稠密模型出现更快的退化？

对三种模型在四种温度配置下生成的 9,360 个结果进行分析后，主要发现是：**决定温度鲁棒性的是指令微调，而非架构稀疏性**。

6.1 主要结论

1. **稀疏架构不会导致温度敏感性**。Mixtral（稀疏、经过指令微调）在所有温度设置下都保持了稳定的性能，可靠性与稠密模型相当或更高。2. **指令微调能够增强鲁棒性**。两种经过指令微调的模型（稀疏和稠密）都表现稳定，而基础模型则出现了退化。这表明，对齐训练能提供一个锚点，使模型即使面对采样随机性，也能保持一致的行为。3. **合规性是非此即彼的**。OLMoE（基础模型）的合规率为 0%；Qwen（经过指令微调）的合规率为 100%。这一差距源于训练方法，而非架构。4.**失效模式具有系统性**。基础模型会因格式不合规而失效；经过指令微调的模型则会出现计算错误。这些是不同的问题，需要采取不同的干预措施。

6.2 实践建议

对于在确定性任务中部署模型的实践者而言，使用经过良好调优的模型时，适中的温度（最高可达 $T = 1.0$）不太可能显著损害可靠性。模型训练质量比稀疏与稠密架构更为重要。

然而，对于要求严格格式合规性（JSON 输出、结构化响应）的应用，仍建议采用较低温度或贪心解码。即使是经过良好调优的模型，在较高温度下也可能出现更多格式违规。

6.3 未来工作

改进后的实验设计将评估析因设计中的全部四种组合： 2×2

- 稀疏模型 + 基础模型
- 稀疏模型 + 指令微调
- 稠密 + 基础模型
- 稠密 + 指令微调

此外，通过匹配激活参数来控制模型规模（例如，3B 稀疏模型与 3B 稠密模型）将增强因果推断。

当前研究结果表明，只要模型经过适当的指令微调，对于 MoE 模型在温度采样下可靠性的担忧可能是不必要的。

代码可用性

所有代码和评估数据均已公开发布，地址：

https://github.com/kabirrgrover/lm-reliability-sparsity

参考文献

[1] William Fedus、Barret Zoph 和 Noam Shazeer。Switch Transformer：通过简单高效的稀疏性扩展至万亿参数模型。《机器学习研究杂志》，2022 年。[2] Ari Holtzman、Jan Buys、Li Du、Maxwell Forbes 和 Yejin Choi。神经文本退化的奇特案例。发表于国际学习表征会议，2020 年。[3] Albert Q. Jiang 等人。Mixtral 专家模型。*arXiv* 预印本 *arXiv:2401.04088*，2024 年。[4] Niklas Muennighoff 等人。OLMoE：开放式混合专家语言模型。*arXiv* 预印本 *arXiv:2409.02060*，2024 年。[5] Noam Shazeer 等人。超大规模神经网络：稀疏门控混合专家层。发表于国际学习表征会议，2017 年。[6]Yizhong Zhou 等人。大语言模型的指令遵循评估。*arXiv* 预印本 *arXiv:2311.07911*，2023 年。