多项选择题回答任务中的保形 P 值及有风险控制证明

Yuanchang Ye
School of Data Sciences
Zhejiang University of Finance & Economics
HangZhou, China
yuanchang0213@zufe.edu.cn

August 15, 2025

ABSTRACT

本研究介绍了一种增强显著性检验的保序预测（CP）框架，以提升大型语言模型（LLMs）在多项选择题回答（MCQA）中的可信度。虽然 LLMs 在专业问答场景中已被越来越多地应用，但幻觉和非事实生成显著影响了响应可靠性。尽管 CP 为预测集提供了统计上严格的边际覆盖保证，显著性检验也提供了已确立的统计严谨性，但它们的协同整合尚未被探索。为减轻幻觉和事实性错误，我们的框架通过自一致性重采样 MCQA 响应整合 p 值计算与保序评分。该方法计算选项频率以应对 LLMs 的黑箱性质，随后通过经验性推导的 p 值进行零假设检验（$\mathcal{H}_0$ ）构建预测集。使用现成的 LLMs 在 MMLU 和 MMLU-Pro 基准上的评估表明：（1）增强的 CP 实现了用户指定的经验误盖率；（2）随着风险水平（α ）的增加，测试集平均预测集大小（APSS）单调减小，验证了 APSS 作为有效不确定性度量。这项工作为在高风险问答应用中部署可信的 LLM 建立了一个有原则的统计框架。

Keywords Large Language Models ; Conformal Prediction ; Significance Test ; Multiple Choice Questions Answering ; Average Prediction Set Size

1 引言

大型语言模型（LLM）的迅速发展使得它们在包括 AI 支持的客户服务、内容生成和知识问答等领域得到了广泛应用 [3, 4, 5, 18] 。尽管自然语言生成能力在不断改进，研究表明 LLM 仍然持久地表现出事实幻觉，自信地生成错误或非事实的文本。这些输出通常展现出逻辑一致性、权威性展示和高说服力，而包含事实偏差或虚构内容。直接使用幻觉内容有造成错误决策、误导用户以及最终破坏高风险应用中可信度的风险 [2, 6, 9] 。因此，开发高效的自动幻觉检测框架构成了确保 LLM 可靠性和应用安全性的关键挑战 [8] 。

不确定性量化（UQ）对于评估模型可靠性、风险管理和幻觉识别是至关重要的 [12, 15, 14] 。由于大型语言模型（LLMs）主要提供基于文本的接口，目前流行的置信度评估方法包括校准技术、口头不确定性表达和启发式方法。然而，这些启发式方法不能提供任务特定的性能保证，限制了其可靠性。虽然保序预测（CP）提供模型无关且统计上严格的不确定性估计，但其在自然语言生成（NLG）中的应用仍然具有挑战性。为了解决这一问题，我们采用统计显著性检验方法，制定严格的数学构造，确保性能保证，同时保持统计鲁棒性。

我们的基于显著性检验的保形预测（ST-CP）方法与传统的 CP 方法本质上不同，[1] ，通过在最小假设下提供统计覆盖保证，同时实现卓越的计算效率。在这项工作中，我们将 ST-CP 应用于多项选择问题回答（MCQA）任务 [17] 。数据根据预定义的比例分为校准集和测试集。在对问题答案进行多次重新采样迭代之后，我们使用统计显著性检验计算 p 值。然后，这些 p 值与预定的显著性水平（α ）进行比较，以决定是否拒绝原假设，从而决定在预测集中包含哪些答案选项。

我们使用 MCQA 数据集 MMLU 和 MMLU_PRO 对我们的方法进行评估，测试了四种大语言模型的变体：Qwen2.5-3B-Instruct、Llama-3.2-3B-Instruct、Meta-Llama-3-8B-Instruct 和 Vicuna-7B-v1.5。为了满足 CP 的可交换性条件，实验在每个特定学科的数据集中进行。广泛的实证结果表明，可以在用户指定的风险水平下精确控制误覆盖率。例如，在风险容忍度较高（$\alpha = 0.5$ ）的 MMLU_PRO 生物学基准上，Vicuna-7B-v1.5 的平均实证错误率保持在 $\alpha = 0.5$ 以下。值得注意的是，无论校准-测试数据集比例如何，该方法始终将实证错

误率限制在预定义的风险水平以下。这证实即使校准数据有限，框架也能保持覆盖保证并表现出强大的鲁棒性。此外，增加风险参数 α 会逐步减小平均预测集大小。这一行为展示了我们方法生成统计严谨、紧凑的预测集的能力，能够有效减少大语言模型输出中的幻觉倾向。

2 相关工作

在 LLMs 中的幻觉：在自然语言处理领域中，幻觉指的是由模型生成的内容，虽然看似连贯，但与源输入（例如提示、上下文）不一致或与可验证的世界知识相矛盾——这一概念来源于心理学研究中的感知失真。在特定任务设置中，如多选问题回答（MCQA），LLM 幻觉主要表现为两种核心类型：内在幻觉（答案直接与问题上下文或选项相冲突）和外在幻觉（答案依赖于不可验证的外部知识而没有明确的上下文矛盾）。鉴于 LLMs 的核心目标是准确遵循指令（例如，选择正确答案），MCQA 幻觉进一步分类为：事实幻觉（偏离客观/问题特定可验证事实）和忠实性幻觉（在语义上未与问题约束保持一致，包括选项误解或逻辑不连贯）。

检测方法主要包括：1）基于外部模型的评估，使用高精度判别器（例如，更高级的 LLM）对回答进行评分——受限于标注质量和可用性；2）基于规则的检测，针对特定维度的幻觉错误（例如，事实/实体不准确），使用 LAMA、TruthfulQA 和 FaithDial 等既定基准进行量化。

然而，大型语言模型在减少幻觉方面依然面临持续的挑战：对比解码等技术通过候选比较或偏好建模来解决认知偏见，但由于干扰项的干扰和对比不稳定性，其在多项选择问答（MCQA）中的效果有限；事后纠正方法（例如，迭代提示）会产生难以承受的计算成本，且缺乏跨任务的普适性。从根本上说，大型语言模型的参数化知识表示和生成机制使得其易受包括知识边界模糊、推理链脆弱和对噪声过于敏感等问题的影响。结果，在要求高精度的 MCQA 任务中，幻觉风险仍然存在。该框架为分析和增强大型语言模型中的内容可靠性提供了系统的方法，特别是针对严格的问答应用。

集成的 SCP 显著性测试框架：拆分一致性预测（SCP）和显著性测试构成了在 LLM 幻觉研究中进行不确定性量化的协同 [16]，统计上严格的范式。SCP 利用可交换的校准数据来构建具有保证覆盖概率（$\geq 1-\alpha$）的预测集，用于黑箱模型输出。其模型无关和分布无关的特性仅要求数据可交换性。对于幻觉检测，SCP 通过置信阈值动态过滤生成的候选项（例如，高置信度错误实体）或使用停止规则截断低似然输出 [13, 10]。

显著性测试通过假设检验识别幻觉现象 [7, 11]（例如，生成概率的 p 值计算），检测与可验证知识/上下文显著偏离的输出。其主要应用包括：1）通过多重检验校正控制错误发现率（FDR）（例如，用于识别幻觉实体的 Benjamini-Hochberg 程序）；2）使用组合检验证多源一致性（例如，费舍尔方法）。

集成框架双向运行：SCP 限定候选空间以简化显著性检验，而显著性结果动态优化 SCP 的置信阈值——通过控制幻觉遗漏率来增强覆盖可靠性。尽管在非交换数据适应、多步误差传播以及计算开销上存在挑战，SCP-显著性检验提供了一个理论上有依据且在实践中可行的框架，用于在大型语言模型（LLMs）中量化幻觉、降低风险和进行可靠性评估，利用了严格的统计保证和模型无关性。

3 方法

我们的方法主要解决两个挑战：（1）如何使用显著性检验和保形预测（CP）在测试数据上构建具有边际保证的预测集，以及（2）如何证明我们的方法满足统计显著性要求。在本节中，我们首先介绍基本符号定义。随后，我们展示基本的保形预测框架，通过该框架我们为显著性检验构建 p 值。最后，我们提供实现边际保证的理论证明。

3.1 初步

我们将这个任务表述为一个有 K 个不同选项的多项选择问答 (MCQA) 问题。遵循传统的 CP 方法论，设 $\mathcal{D}_{\text{cal}} = \{(x_i, y_i^*)\}_{i=1}^n$ 表示包含真实标签的校准数据集，其中 n 表示校准示例的总数量，y_i^* 表示 i 个实例的真实标签。我们考虑一个用于预测的测试输入输出对 (x_{n+1}, y_{n+1})。显著性水平 $\alpha \in (0,1)$ 具有双重目的：确定假设检验中的 Typel 错误率和指定 CP 中的期望覆盖率。我们将大语言模型 (LLM) 视为一个黑盒预测器 $\hat{f} : \mathcal{X} \to \mathcal{Y}$，将输入特征映射到输出标签，而对其内部机制保持不可知论。CP 框架中的一个关键组成部分是非符合性分数 $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$，该分数试探性地衡量输入与候选标签之间的兼容性。对于分类任务，我们定义 $\mathcal{S}(x, y) = 1 - \hat{f}(y|x)$，其中 $\hat{f}(y|x)$ 表示模型对类别 y 的估计概率，$S_i = \mathcal{S}(x_i, y_i)$ 表示 i 个校准示例的非符合性分数。

对于每个校准例子 (x_i, y_i^*)，我们从 LLM 进行 P 次独立采样以获得响应集 $\{\hat{y}_1^{(i)}, \hat{y}_2^{(i)}, \hat{y}_3^{(i)}, \ldots, \hat{y}_K^{(i)}\}$。我们计算真实标签出现的经验频率为 $\hat{f}(y|x)$，随后计算非一致性分数 $S(x_i, y^*)$。对于具有多个有效答案的 MCQA 问题，$\hat{f}(y|x)$ 对应于所有正确选项的总频率。对于预测未见测试实例的置信预测集的生成过程 x_{test}，包含三个阶段：

Algorithm 1 生成预测集

Input: $(x_i, y_i^*)_{i=1}^n$, (x_{test}, y) , α
Output: prediction set $\mathcal{C}_\alpha(x_{test})$

// Construct calibration data
Initialize $\mathcal{D}_{\text{cal}} = \{(x_i, y_i^*)\}_{i=1}^n$
for each $x_i \in \mathcal{D}$ **do**
  $s_i \leftarrow 1 - \hat{f}(x_i, y^*)$
end for
sort $\{s_i\}_{i=1}^N$
for each $y_k^{test} \in \mathcal{Y}$ **do**
  calculate $S(x_{test}, y_k^{test})$
  if $\mathbb{P}\left(\frac{\sum_{i=1}^n \mathbf{1}\{s_i > s_{\text{test}}\}+1}{n+1} \leq \alpha\right) \leq \alpha$ **then**
    continue
  else
    $\mathcal{C}_\alpha(x_{test}) \leftarrow S(x_{test}, y_k^{test})$
  end if
end for
return Prediction Set $\mathcal{C}_\alpha(x_{test})$

1. 计算校准数据中的不符合度分数 $\{s_1, \ldots, s_n\}$ ，其中 $s_i = S(x_i, y_i^*)$
2. 从经验分数分布中确定保形 α 分位数 $\tau = Q_{1-\alpha}(\{s_i\}_{i=1}^n)$ 。根据标准 CP 实践，我们计算 τ 为排序后的校准分数中第 $[(1-\alpha)(n+1)]$ 小的值
3. 构建预测集：

$$C_\alpha(x_{\text{test}}) = \left\{ y : S(x_{\text{test}}, y) \leq \text{Q}\left(\{s_i\}_{i=1}^n, \frac{[(n+1)(1-\alpha)]}{n} \right) \right\} \tag{1}$$

这保证了：

$$\mathbb{P}\left(y_{\text{test}}^* \in C_\alpha(x_{\text{test}})\right) \geq 1 - \alpha \tag{2}$$

步骤 1-2 构成了校准阶段，而步骤 3 生成最终预测集。直观地，该集合包括了所有与测试输入的兼容性超过由校准分布建立的 α 分位数阈值的标签。

3.2 显著性检验

根据统计显著性检验的基本原则，我们建立了一个 p 值，用于与我们预定义的显著性水平 α 进行比较。通过使用经验累积分布函数（ECDF）重新表述公式（1），我们得出：

$$\mathcal{C}_\alpha(x_{\text{test}}) = \left\{ y : \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{s_i \leq S(x_i, y)\} \leq \frac{\lceil (n+1)(1-\alpha) \rceil}{n} \right\} \tag{3}$$

其中，$\mathbf{1}\{\cdot\}$ 表示指示函数。这个公式意味着，如果校准得分小于其不一致性得分 $S(x_{\text{test}}, y)$ 的比例不超过调整后的阈值 $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ ，则包含一个标签 y 。通过对方程（3）的代数操作，我们得到等效条件：

$$\frac{\sum_{i=1}^n \mathbf{1}\{s_i > S(x_{\text{test}}, y)\} + 1}{n+1} > \alpha \tag{4}$$

这种等价性建立了我们的假设检验框架：

- 零假设 $\mathcal{H}_0$: y 是 x_{test} 的真实标签
- 备择假设 $\mathcal{H}_1$: y 不是真实标签

我们定义 p 值统计量为：

$$p(y) = \frac{\sum_{i=1}^n \mathbf{1}\{s_i > S(x_{\text{test}}, y)\} + 1}{n+1} \tag{5}$$

Table 1: 不同数据集和模型在分割比例为 0.5 下的平均错误率

Dataset	Model	Alpha Rate								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
MMLU	Llama-3.1-8B-Instruct	0.088	0.187	0.288	0.389	0.488	0.584	0.684	0.785	0.877
	Llama-3.2-3B-Instruct	0.090	0.187	0.284	0.388	0.487	0.582	0.681	0.782	0.877
	Meta-Llama-3-8B-Instruct	0.091	0.192	0.288	0.392	0.490	0.585	0.686	0.785	0.876
	Qwen2.5-3B-Instruct	0.091	0.190	0.290	0.392	0.488	0.584	0.683	0.783	0.879
	Qwen2.5-7B-Instruct	0.090	0.189	0.289	0.389	0.486	0.582	0.683	0.785	0.877
	openchat_3.5	0.088	0.186	0.284	0.386	0.483	0.579	0.679	0.782	0.876
	vicuna-7b-v1.5	0.091	0.191	0.289	0.391	0.490	0.582	0.684	0.782	0.875
MMLU-PRO	Llama-3.1-8B-Instruct	0.099	0.197	0.296	0.395	0.494	0.596	0.693	0.793	0.893
	Llama-3.2-3B-Instruct	0.098	0.197	0.297	0.397	0.496	0.597	0.697	0.796	0.893
	Meta-Llama-3-8B-Instruct	0.097	0.197	0.297	0.397	0.498	0.597	0.697	0.796	0.895
	Qwen2.5-3B-Instruct	0.097	0.196	0.296	0.397	0.496	0.595	0.695	0.794	0.894
	Qwen2.5-7B-Instruct	0.097	0.196	0.295	0.395	0.495	0.595	0.696	0.794	0.894
	openchat_3.5	0.099	0.199	0.298	0.396	0.497	0.596	0.695	0.796	0.894
	vicuna-7b-v1.5	0.098	0.199	0.298	0.398	0.497	0.596	0.696	0.795	0.893

对于真实标签 y^*_{test}，这可以简化为：

$$p(y^*_{\text{test}}) = \frac{\sum_{i=1}^{n} \mathbf{1}\{s_i > s_{\text{test}}\} + 1}{n+1}, \quad s_{\text{test}} = S(x_{\text{test}}, y^*_{\text{test}}) \tag{6}$$

决策规则遵循标准显著性检验原则：当 $p(y) \leq \alpha$ 时，我们拒绝 $\mathcal{H}_0$，从而将 y 排除在预测集之外。覆盖保证是由以下概率界限产生的：

$$\frac{\sum_{i=1}^{N} \mathbf{1}\{s_i > s_{N+1}\} + 1}{N+1} \leq \alpha \tag{7}$$

方程 (7) 和 (2) 之间的等价性表明，我们的方法通过其与符合性预测理论的联系满足边际覆盖保证。这确立了在标准可交换性假设下我们框架的统计有效性。我们的实验使用 MMLU 及其增强版本 MMLU-Pro 作为主要多项选择题回答 (MCQA) 基准。MMLU 包括 15,908 个问题，涵盖 57 个主题（包括 STEM、人文、社会科学），分为小样本开发集（285 个问题）、验证集（1,540 个问题）和测试集（14,079 个问题），用于评估零样本/小样本知识能力。MMLU-Pro 作为高难度基准，包含 12,032 个问题，其中 5,222 个问题从三个高质量来源（STEM 网站/TheoremQA/SciBench）精选，以增强在九个核心学科的覆盖，尤其是物理（+888）、化学（+954）和工程（+902）。相比 MMLU 的四选项设计，该基准的创新十选项设计显著提高了区分难度。这些基准集合通过其学科多样性、高级推理要求和严格的划分协议，为幻觉检测和知识边界评估提供了严格的评估平台。

3.2.1 大型语言模型

我们使用四个顶尖的语言模型进行全面评估：Qwen2.5-3B-Instruct、Llama-3-3B-Instruct、Meta-Llama-3-8B-Instruct 和 Vicuna-7B-v1.5，并进行后续性能可视化。针对参数变体进行特定架构的对比分析。主要配置包括：

- Qwen2.5-3B-Instruct：特征为分组查询注意力（GQA）和 SwiGLU 激活，以实现高效解码。采用三级对齐流程（多语言预训练 监督微调 RLHF 优化），优化用于 128K 上下文的复杂指令推理。
- Llama-3-3B-Instruct：整合了 KV 缓存和 RMSNorm 标准化。通过双阶段范式（预训练 + PPO-RLHF 对齐）训练，以实现高效的 8K 上下文多语言处理。
- Meta-Llama-3-8B-Instruct：通过扩展隐藏维度（6,656）增强表示能力。结合 DPO-RLHF 优化，在 8B 参数类别中实现了最先进的推理性能。
- Vicuna-7B-v1.5：仅在会话数据（125K 多轮对话）上进行微调，在 4K 上下文交互中表现出强大的连贯性。

3.2.2 实现细节

通过综合显著性测试和分割保序预测（SCP）实施 MCQA 预测集的边缘覆盖保障。方法流程包括：(1) 回复采样：每个问题进行 20 次独立生成（温度 =1.0，top-p=0.9），输出长度限制为单个 token 以进行多选题不确定性量化。(2) 数据划分：基于主题的数据集通过 SCP 按预定比例分为校准集和测试集，选项概率由模型输

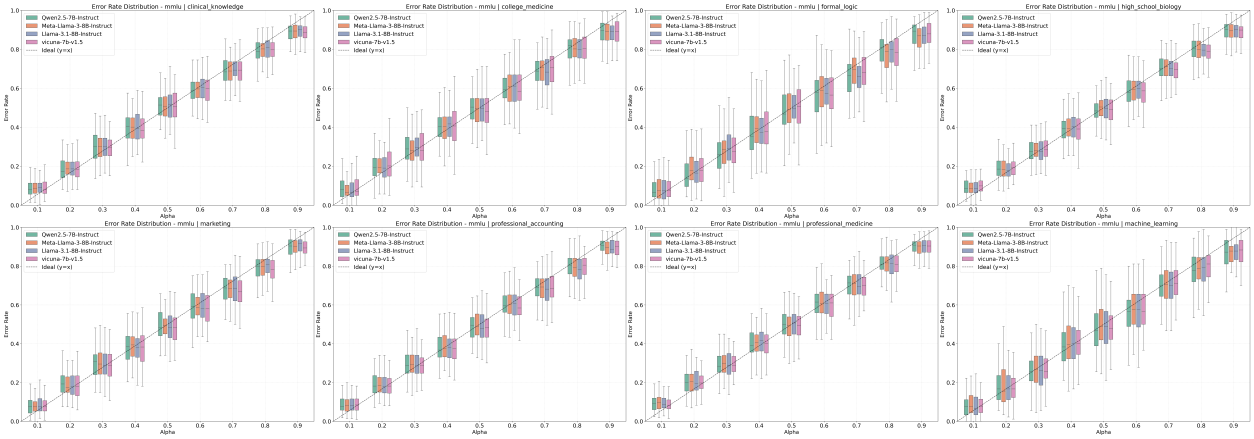

Figure 1: 来自 MMLU 基准测试中 8 名被试的误差率分布的箱线图，具有较小的四分位距

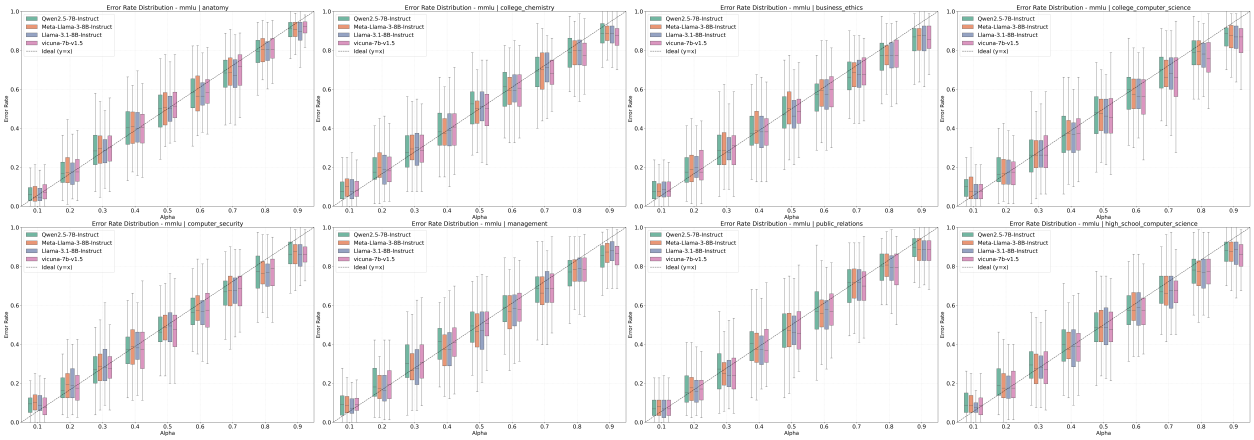

Figure 2: 从 MMLU 基准中选择的具有较高四分位距的 8 个主题的错误率分布箱线图

出推导。(3) 统计检验：从校准数据中计算不一致分数。显著性测试通过为测试集选项确定 p 值，当 p 值低于阈值 α 时拒绝原假设（选项排除标准）。(4) 评估指标：根据真实标签计算经验错误率。通过 100 次自举重采样生成箱线图量化预测集的可靠性，同时通过不同显著性水平的预测集大小分布提供边缘覆盖保障的补充评估。

本节严格验证了通过方程（3）构建的校准预测集，确认其能够可靠地实现用户指定的错覆盖率目标覆盖水平。我们通过不确定性度量引导的选择性预测进一步考察了该框架的实际效用。

对名义覆盖水平的实证验证是通过将每个数据集以 1:1 的比例分为校准集和测试集进行的。共形不确定性阈值根据预先指定的错误覆盖率从校准数据中得出。随后对测试数据进行的覆盖率评估产生了图 2-4 中显示的 MMLU 和 MMLU-Pro 结果。

首先，我们将数据的分割比率设置为 0.5。我们通过 100 次采样测量了每个数据集中不同对象的准确性。最后，我们汇总并平均了所有对象的准确性，并将范围扩展到拥有更多参数的模型。我们得到以下结果：在不同的数据集和模型下，预测集的错误率低于我们设定的显著性水平。这表明，在我们相对标准的分割比率下，获得的结果符合边际保证。此外，所有结果并没有显著低于我们设定的显著性水平，而是在一个小范围内低于设定值。这确保了不会超过该方法的上限，也不会低于下限。

图 2 分析了八个最小内距四分位数（IQR）的 MMLU 子任务在不同置信水平上的错误率分布。每个子任务的可视化图将置信参数与错误率进行对比，使用箱线图描绘四个 LLM 变体的四分位距、中位数和数据范围，并用来表明置信参数与错误率的理论等价性作为参考。

所有模型的中位数误差率随着 α 增加而单调递减，表明与 α 误差率呈反相关。IQR 分析显示，在极端 α 值（例如 0.1, 0.9）时变异性降低，但在中间值（例如 0.4-0.6）时分散性显著增加，表明在高置信度预测下误差

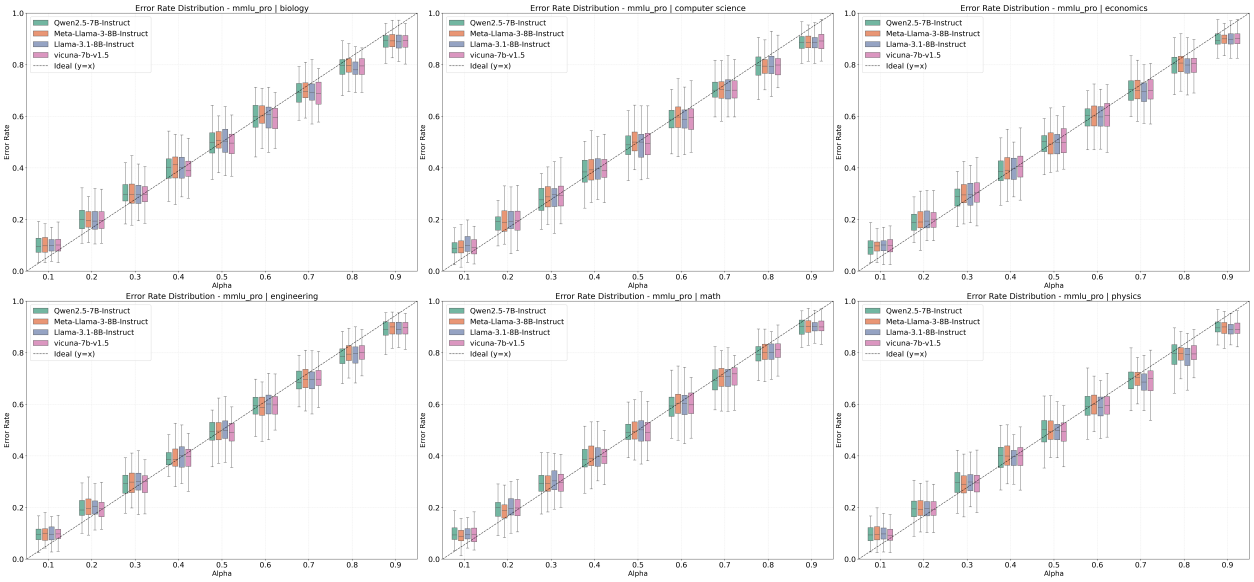

Figure 3: 来自 MMLU-PRO 基准测试的 6 名受试者的错误率分布箱形图

Table 2: 在不同风险水平下预测集大小的结果

Dataset	LLMs/ α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
MMLU	Llama-3.1-8B-Instruct	1.98	1.48	1.16	0.89	0.66	0.49	0.34	0.22	0.11
	Meta-Llama-3-8B-Instruct	2.07	1.54	1.19	0.92	0.69	0.50	0.35	0.22	0.11
	Qwen2.5-7B-Instruct	1.98	1.41	1.07	0.83	0.64	0.47	0.33	0.22	0.11
	vicuna-7b-v1.5	2.94	2.32	1.89	1.50	1.12	0.82	0.55	0.33	0.14
MMLU-PRO	Llama-3.1-8B-Instruct	6.00	4.51	3.33	2.44	1.73	1.20	0.78	0.44	0.21
	Meta-Llama-3-8B-Instruct	6.78	5.12	3.82	2.78	1.95	1.28	0.79	0.44	0.20
	Qwen2.5-7B-Instruct	6.58	5.01	3.61	2.40	1.60	1.04	0.64	0.37	0.18
	vicuna-7b-v1.5	7.86	6.70	5.53	4.43	3.39	2.44	1.62	0.92	0.41

率的一致性更好。最值得注意的是，Qwen2.5-3B-Instruct 表现最为出色：它实现了最低的中位数误差率，约 75 % 的观测值低于其目标 alpha 值，这表明在评估的子任务中准确性和稳定性得到了增强。

箱线图展示了八个高离散度 MMLU 科目的错误率分布，揭示了与最小 IQR 情况下观察到的相似的经验错误率趋势。然而，这些科目显示出显著更大的变异性。对每个 α 水平上的取值范围进行比较分析显示，实验试验中的错误率波动更大，表明模型的鲁棒性较低。然而，Vicuna-7B-v1.5 在这些特定科目任务中表现优异，表现为较低的中位数错误率和较小的四分位距。

在 MMLU-Pro 中，由于不同模型之间的结果高度相似，因此展示了六个具有代表性的主题。尽管选项复杂性和难度增加，模型仍然保持稳定的经验误差率。在极值和中间 α 值之间的四分位距差异不太明显。经验误差率随着 α 的增加而持续下降，而在相同 α 水平下的值范围显著减少，这表明数据集的稳定性得到了增强。除了 Qwen2.5-3B-Instruct 和 Vicuna-7b-v1.5 的持续表现外，Llama-3.1-8B-Instruct 的中位误差率也达到了与领先模型相当的水平。

这项研究从两个不同的角度评估了模型对预测集大小的影响：一个是个体对象的集合大小，另一个是整个数据集的平均集合大小。两个指标都在两个独立的数据集上进行了测试。

表格数据证实，增加置信度阈值 α 会单调减少所有模型的平均预测集合大小（APSS），这与选择性预测理论一致。Llama 系列模型展示了卓越的压缩效率，特别是在高置信度区域（$\alpha \geq 0.7$ ），在 $\alpha = 0.8$ 处比 Vicuna 模型实现了 33 % 更小的预测集合。相比之下，Qwen2.5 模型表现出逐步减少的模式，在中等置信度水平（$\alpha = 0.5$ ）保持较大的预测集合。Vicuna 模型始终产生最高的 APSS 值，表明预测不确定性显著提高。

数据集复杂性系统性地调整不确定性管理。在更具挑战性的基准中，所有模型的预测集大小均超过基线尺寸 3 倍以上。这一差异在较高的置信水平下进一步放大，最大模型间差异达到 ×，置信水平为 $\alpha = 0.6$ 。值得注意的是，Qwen2.5 在严格的条件下（$\alpha = 0.9$ ）表现出相对优势，表明其在高不确定性环境中具有更强的覆盖稳定性。这种任务依赖行为需要特定背景下的校准。

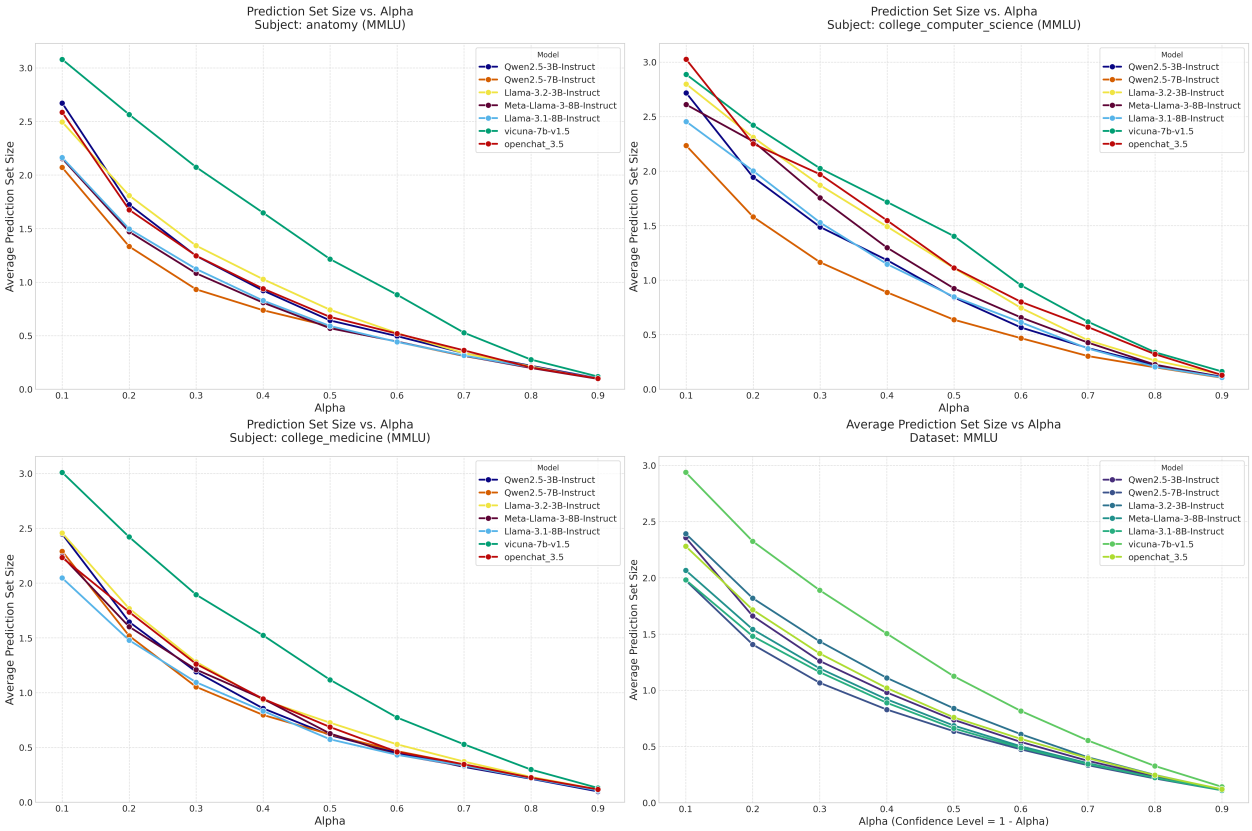

Figure 4: 个体受试者与 MMLU 中平均预测集合大小的比较曲线

APSS 指标有效量化了不确定性管理的效率。Llama 的出色压缩反映了精确的置信度校准，非常适合效率关键的应用。Qwen2.5 的渐进式衰减为对可靠性敏感的环境提供了更稳定的风险覆盖。这些发现确立了 APSS 作为诊断工具的地位，可以揭示校准的局限性，并通过平衡压缩-覆盖策略为风险分层决策场景提供模型优化的指导。

在整个 MMLU 数据集和各个科目中，预测集大小的变化揭示了三种主要模式（图 5）。随着显著性水平 α 的增加，预测集大小单调递减，这表明 α 在置信度-可靠性权衡中的调节作用：较高的 α 值产生更高置信度的预测，但候选类别更少，而较低的 α 保留更多的类别以确保可靠性。在 $\alpha = 0.4$，大多数模型收敛到基数为 1 的预测集，这将这一阈值确立为实践中 α 配置的效率与准确性之间的关键平衡点。观测到模型之间存在显著的行为差异：Vicuna-7B-v1.5 在低 α（$\alpha < 0.2$）时保持较大的初始集大小，并近似线性地减少，表明其保守行为适用于高准确度关键的应用，而其他模型在 $\alpha = 0.1$ 和 $\alpha = 0.2$ 之间快速缩小集大小，随后衰减率稳定，更适合效率敏感的场景。总体而言，这些结果勾勒出 α 对预测集基数的系统控制，而特定模型的响应特征则为情境优化的模型选择提供了信息。

与 MMLU 数据集类似，MMLU-Pro 在 α 增加时，预测集合大小呈递减趋势。然而，在 $\alpha = 0.1$ 和 $\alpha = 0.2$ 之间的减少不太明显。在整个 $\alpha \in [0.1, 0.5]$ 期间，集合大小近似线性减少，当 α 值较高时，减少率降低。对个体模型的分析显示，Vicuna-7B-v1.5 始终保持最大的预测集合（表示更大的预测差异），而其他模型显示出相似的性能，并提供一致的跨科目结果，证实了我们预测集合计算方法的稳定性。

总体而言，这两个数据集证实了预测集大小随着 α 的增加而单调收缩。与第 4.2 节的结果相关联，这种收缩意味着覆盖率不足的风险提高。然而，在接近 $\alpha = 0.4$ 时，模型会收敛到单一的预测集合，同时保持较低的经验误差率。这支持在 $\alpha \approx 0.4$ 处平衡覆盖不足风险和集合大小，以实现最佳选择性预测性能。

4 结论

本研究引入了一种增强的保形预测（CP）框架，它结合了统计显著性测试以实现领域特定多项选择题回答（MCQA）任务中的风险控制。所提出的框架提供了统计严格且可解释的方法论，可以在减轻模型幻觉风险的同时，增强大型语言模型（LLMs）在学科应用中的可信度。我们的方法在保留 CP 的模型不可知和分布无

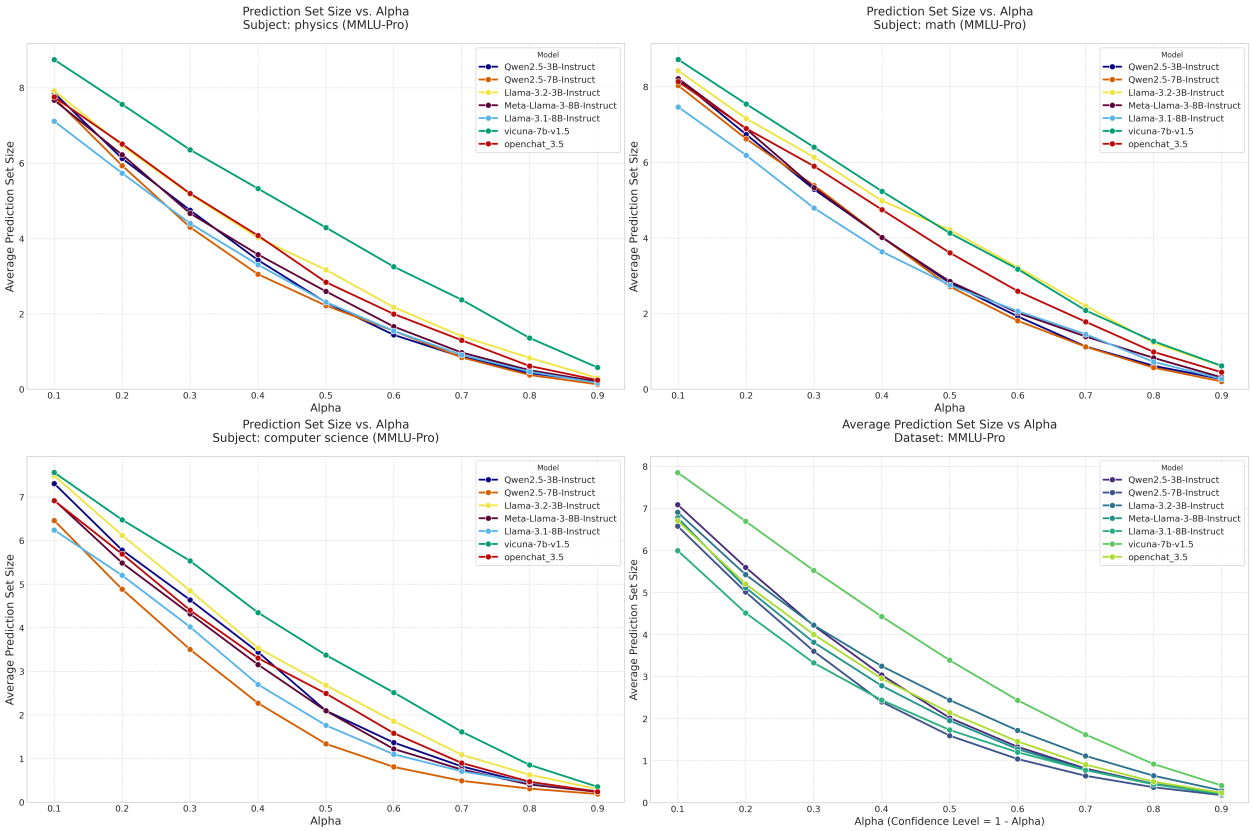

Figure 5: 个体主题与平均预测集大小在 MMLU-pro 中的比较曲线

关优势的同时，简化了计算过程，并将功能扩展到复杂风险管理场景的基础覆盖控制之外。关键是，该框架始终实现用户指定的错误覆盖率（α）。对显著性水平（α）及相应集合大小的预测集合错误率分析显示，平均预测集大小（APSS）与风险水平之间存在明确的反相关关系，这表明 APSS 在量化 LLM 不确定性方面的实用性。未来的工作将探讨适应其他专业 QA 领域及与先进 LLM 架构的整合以进一步提高可靠性和透明度。

References

[1] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[2] Jinhe Bi, Yifan Wang, Danqi Yan, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *arXiv preprint arXiv:2502.12119*, 2025.

[3] Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. LLaVA steering: Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.

[4] Jinhe Bi, Danqi Yan, Yifan Wang, Wenke Huang, Haokun Chen, Guancheng Wan, Mang Ye, Xun Xiao, Hinrich Schuetze, Volker Tresp, et al. Cot-kinetics: A theoretical modeling assessing lrm reasoning process. *arXiv preprint arXiv:2505.13408*, 2025.

[5] Haokun Chen, Hang Li, Yao Zhang, Jinhe Bi, Gengyuan Zhang, Yueqi Zhang, Philip Torr, Jindong Gu, Denis Krompass, and Volker Tresp. Fedbip: Heterogeneous one-shot federated learning with personalized latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[6] Haokun Chen, Yueqi Zhang, Yuan Bi, Yao Zhang, Tong Liu, Jinhe Bi, Jian Lan, Jindong Gu, Claudia Grosser, Denis Krompass, et al. Does machine unlearning truly remove model knowledge? a framework for auditing unlearning in llms. *arXiv preprint arXiv:2505.23270*, 2025.

[7] Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. *Advances in Neural Information Processing Systems*, 37:73884–73919, 2024.

[8] Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. Do llms know when to not answer? investigating abstention abilities of large language models. *arXiv preprint arXiv:2407.16221*, 2024.

[9] Xuankun Rong, Wenke Huang, Jian Liang, Jinhe Bi, Xun Xiao, Yiming Li, Bo Du, and Mang Ye. Backdoor cleaning without external guidance in mllm fine-tuning. *arXiv preprint arXiv:2505.16916*, 2025.

[10] Qingni Wang, Tiantian Geng, Zhiyuan Wang, Teng Wang, Bo Fu, and Feng Zheng. Sample then identify: A general framework for risk control and assessment in multimodal large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[11] Xiaoning Wang, Yuyang Huo, Liuhua Peng, and Changliang Zou. Conformalized multiple testing after data-dependent selection. *Advances in Neural Information Processing Systems*, 2024.

[12] Yujun Wang, Jinhe Bi, Yunpu Ma, and Soeren Pirk. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. *arXiv preprint arXiv:2506.14766*, 2025.

[13] Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Heng Tao Shen, and Xiaofeng Zhu. Conu: Conformal uncertainty in large language models with correctness coverage guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

[14] Zhiyuan Wang, Jinhao Duan, Qingni Wang, Xiaofeng Zhu, Tianlong Chen, Xiaoshuang Shi, and Kaidi Xu. Coin: Uncertainty-guarding selective question answering for foundation models with provable risk guarantees. *arXiv preprint arXiv:2506.20178*, 2025.

[15] Zhiyuan Wang, Jinhao Duan, Chenxi Yuan, Qingyu Chen, Tianlong Chen, Yue Zhang, Ren Wang, Xiaoshuang Shi, and Kaidi Xu. Word-sequence entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond. *Engineering Applications of Artificial Intelligence*, 2025.

[16] Zhiyuan Wang, Qingni Wang, Yue Zhang, Tianlong Chen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. SConU: Selective conformal uncertainty in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.

[17] Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*, 2024.

[18] Gengyuan Zhang, Jinhe Bi, Jindong Gu, Yanyu Chen, and Volker Tresp. Spot! revisiting video-language models for event understanding. *arXiv preprint arXiv:2311.12919*, 2023.

A 证明

Proof. 我们通过以下步骤建立 p 值公式与保形预测（CP）之间的等价性：步骤 1：保形预测集的构建
对于测试点 x_{N+1} ， 显著性水平 α 下的保形预测集为：

$$\mathcal{C}_\alpha(x_{N+1}) = \left\{ y : S(x_{N+1}, y) \leq Q_{1-\alpha}^{(N)} \right\} \tag{8}$$

其中 $Q_{1-\alpha}^{(N)} = \text{Quantile}\left(\{s_i\}_{i=1}^N, \frac{\lceil (N+1)(1-\alpha) \rceil}{N} \right)$ 是从校准集中得出的符合性得分 $\{s_i = S(x_i, y_i^*)\}_{i=1}^N$ 的经验分位数。这保证了：

$$\mathbb{P}\left(y_{N+1}^* \in \mathcal{C}_\alpha(x_{N+1}) \right) \geq 1 - \alpha \tag{9}$$

在校准数据 $\{(x_i, y_i)\}_{i=1}^N$ 和测试点 (x_{N+1}, y_{N+1}^*) 上具有概率。步骤 2：p 变量的定义和等价性
为候选标签 y 定义 p 变量：

$$P(y) = \frac{1}{N+1} \left(1 + \sum_{i=1}^{N} \mathbb{I}\{ s_i > S(x_{N+1}, y) \} \right) \tag{10}$$

对于真实标签 $y = y_{N+1}^*$ ， 我们有：

$$P(y_{N+1}^*) = \frac{1}{N+1} \left(1 + \sum_{i=1}^{N} \mathbb{I}\{ s_i > s_{N+1} \} \right) \tag{11}$$

其中 $s_{N+1} = S(x_{N+1}, y^*_{N+1})$ 。步骤 3：覆盖等价
包含条件 $y \in \mathcal{C}_\alpha(x_{N+1})$ 等同于：

$$S(x_{N+1}, y) \leq Q^{(N)}_{1-\alpha} \tag{12}$$

$$\iff \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\{s_i \leq S(x_{N+1}, y)\} \leq \frac{\lceil (N+1)(1-\alpha)\rceil}{N} \tag{13}$$

$$\iff \sum_{i=1}^{N}\mathbb{I}\{s_i \leq S(x_{N+1}, y)\} < \lceil (N+1)(1-\alpha)\rceil \tag{14}$$

$$\iff \sum_{i=1}^{N}\mathbb{I}\{s_i > S(x_{N+1}, y)\} > N - \lceil (N+1)(1-\alpha)\rceil \tag{15}$$

$$\iff \frac{1}{N+1}\left(1 + \sum_{i=1}^{N}\mathbb{I}\{s_i > S(x_{N+1}, y)\}\right) > \frac{N - \lceil (N+1)(1-\alpha)\rceil + 1}{N+1} \tag{16}$$

$$\iff P(y) > \alpha \quad \text{(for non-integer boundaries)} \tag{17}$$

因此，$y \in \mathcal{C}_\alpha(x_{N+1}) \iff P(y) > \alpha$ 。步骤 4：I 型错误控制
对于真实标签 y^*_{N+1} ，我们有：

$$\mathbb{P}\left(P(y^*_{N+1}) \leq \alpha\right) = \mathbb{P}\left(\frac{1}{N+1}\left(1 + \sum_{i=1}^{N}\mathbb{I}\{s_i > s_{N+1}\}\right) \leq \alpha\right) \tag{18}$$

$$= 1 - \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\{s_i \leq s_{N+1}\} < \frac{\lceil (N+1)(1-\alpha)\rceil}{N}\right) \tag{19}$$

$$\leq \alpha \quad \text{(by conformal coverage guarantee)} \tag{20}$$

这满足了对所有 $\alpha \in (0, 1)$ 的 p 变量条件 $\mathbb{P}(P \leq \alpha) \leq \alpha$ 。步骤 5：假设检验等价
考虑假设检验：

$$\mathcal{H}_0 : \quad y = y^*_{N+1} \quad \text{(true label)}$$
$$\text{Reject } \mathcal{H}_0 \text{ if} \quad P(y) \leq \alpha$$

预测集正好是：

$$\mathcal{C}_\alpha(x_{N+1}) = \{y : P(y) > \alpha\} = \{y : \text{fail to reject } \mathcal{H}_0 \text{ for } y\} \tag{21}$$

误覆盖率为：

$$\mathbb{P}(y^*_{N+1} \notin \mathcal{C}_\alpha(x_{N+1})) = \mathbb{P}(P(y^*_{N+1}) \leq \alpha) \leq \alpha \tag{22}$$

因此，保形预测集的构建等价于使用定义的 p 变量的假设检验框架，具有相同的覆盖保证和错误率控制。 □