使用大型语言模型进行价格预测的分位数回归

Nikhita Vedula[1]* **Dushyanta Dhyani**[1]* **Laleh Jalali**[1]

Boris Oreshkin[1] **Mohsen Bayati**[1,2] **Shervin Malmasi**[1]

[1] Amazon.com, Inc. [2] Stanford University
{ veduln, dhyanidd, lalehjal, oreshkin, bayatim, malmasi } @amazon.com

Abstract

大型语言模型（LLMs）在结构化预测任务中表现出潜力，包括回归，但现有方法主要集中于点估计，缺乏对不同方法的系统比较。我们研究使用 LLMs 进行非结构化输入的概率回归，解决具有挑战性的文本到分布预测任务，例如价格估算，其中文本理解的细微差别和不确定性量化至关重要。我们提出了一种新的分位数回归方法，使 LLMs 能够生成完整的预测分布，优于传统的点估计。通过在三个不同的价格预测数据集上的广泛实验，我们证明了一个经过分位数头细化的 Mistral-7B 模型在点和分布估计方面显著优于传统方法，测量标准包括预测准确性和分布校准的三种已建立指标。我们系统地比较了 LLM 方法、模型架构、训练方法和数据扩展，发现 Mistral-7B 在编码器架构、基于嵌入的方法和少样本学习方法上均表现优异。我们的实验还揭示了 LLM 辅助标签校正在实现无系统偏差的人类水平准确性方面的有效性。我们的精心制作的数据集已开放 *，以支持未来的研究。

1 介绍

大型语言模型（LLMs）在广泛的任务中展示了卓越的能力，包括非结构化文档处理 (Zou et al., 2025)、遵循指令 (Ouyang et al., 2022)、多模态推理以及通用智能基准测试 (OpenAI et al., 2023; Bubeck et al., 2023)。超越其最初文本生成 (Brown et al., 2020) 的目的，它们最近已被扩展至结构化数值预测任务，如时间序列预测 (Das et al., 2024)。最近的研究显示，它们在回归任务 (Garg et al., 2022; Vacareanu et al., 2024a) 中也表现出色，当通过上下文例子进行提示时，它们能够以令人惊讶的高精度逼近数字映射。

LLMs 与回归的交集对于长期存在的文本回归任务尤为重要，在该任务中，必须将非结构化语言可靠地映射到数值输出 (Bitvai and Cohn, 2015)。传统的回归模型往往在非结构化文本中关键信息所在的应用中表现不佳，如产品描述或财务报告 (Zhang et al., 2024; Gu et al., 2024)，这些应用需要丰富的文本理解。在像产品定价这样的领域，这一点尤为重要，因为类别之间存在异构特征（例如，电视的屏幕技术与汽车的里程），使得传统的统一特征表示不足以捕捉类别特定的动态。

现有使用大型语言模型（LLM）进行回归的研究已经探索了三种主要方法来应对这些挑战：为特定数值预测任务微调 LLM (Morgan and Jacobs, 2024)，使用 LLM 嵌入作为下游回归模型的特征 (Imperial, 2021; Tang et al., 2024)，以及利用上下文学习实现零样本或少样本数值估计 (Vacareanu et al., 2024a)。然而，除了少数几种方法 (Gruver et al., 2023; Qiu et al., 2024) 外，这些方法大多专注于点估计，忽视了一个关键的限制：无法量化不确定性。许多现实世界的应用，如价格预测、需求预测、金融风险评估和医疗保健运营，内在地需要概率输出而非单值预测 (Arora et al., 2023; Qiu et al., 2024; Gürlek et al., 2024)。在这些应用中，概率建模是捕获不确定性和正常变异、缓解风险以及改善决策至关重要的 (Gu et al., 2024)。目前的研究尚未深入探讨使用 LLM 进行概率回归，也未在单一研究中评估不同 LLM 回归方法之间的权衡。

本文首次研究使用 LLMs 对非结构化文本输入进行概率回归，并迈出了系统研究基于 LLM 的回归方法的一步。我们将研究的重点放在价格预测上，这是一个既需要细致解读自由形式文本输入又需要精确分布估计的任务。在金融领域中，理解完整的价格分布是至关重要的，因为准确地建模尾部行为对有效的风险管理至关重要。

总之，本文做出了三个关键贡献。首先，我们提出了一种新颖的基于 LLM 的分位数回归方法，该方法能够生成具有强校准性能的完整分布，同时保持尖锐的预测区间，并在点估计精度上优于传统方法。定性分析表明，我们的模型能够生成适应数据集中不同价格范围和

* Equal contribution.

* https://github.com/vnik18/llm-price-quantile-reg/

不确定因素的良好校准的分布，对于标准化产品产生更紧凑的分布，而对于价格变化较大的项目则产生适当更宽的分布。其次，我们系统地比较了不同的 LLM 架构（仅解码器与仅编码器与文本嵌入上的传统机器学习与上下文学习）、多种损失函数（二次误差与 pinball），以及各种数据规模，同时研究了训练数据污染。我们证明，微调解码器模型（例如，Mistral-7B）优于其他方法。我们的结果证实，模型大小、数据规模和干净的训练集在稳健、可推广的基于 LLM 的概率回归中都扮演着关键角色。第三，我们研究了 LLM 在价格估算任务中的数据清理应用。我们证明，LLM 指导的标签修正可以达到与人工标注相当的准确性而不会引入系统偏差，并且我们提供了三个策划的数据集（亚马逊产品、Craigslist 二手车和二手船），这些数据集具有标准化的数据划分。

2 相关工作

分布、分位数与文本回归： 由 Koenker and Bassett (1978) 引入的分位数回归，超越了传统的点预测方法，通过在不同概率水平 (Kneib et al., 2023) 上估计条件分位数来刻画目标变量的整个条件分布。与最小二乘回归最小化平方误差不同，分位数回归采用分球损失，使其对异常值具有鲁棒性，并能够捕捉跨分布的异质效应。这种方法在医疗保健操作、金融和经济学中尤其有价值，如当建模完整分布有助于捕获典型和极端估计时。文本回归是一种自然语言处理任务，涉及从非结构化文本输入中预测连续的数值，应用于金融预测、选举预测和票房收入估计等领域 (Bitvai and Cohn, 2015; Dereli and Saraclar, 2019) 。

近日的研究开始探索大语言模型（LLM）在分布预测任务中的能力。Gruver et al. (2023) 通过将 LLM 的标记预测映射到连续分布上展示了概率预测，Qiu et al. (2024) 则开发了一种经过微调的 LLM，用于输出能量预测的离散化概率范围。然而，这些方法主要集中于结构化数值序列，而不是从非结构化文本输入中推导分布，而是依赖于带有特殊数字格式的零样本提示或使用预定义输出范围的领域特定微调。我们的研究解决了通过分位数回归直接从非结构化文本预测完整概率分布的更广泛挑战。我们探索并比较了多个文本到分布模型的方法：（1）计算 LLM 嵌入，然后将它们输入到一个独立的分位数预测模型中，（2）提取嵌入并使用训练数据中“邻近”嵌入的结果来近似分布，以及（3）我们提出的将多分位数头直接附加到 LLM 的最后一个隐藏层上，并通过平滑的弹球损失对整个架构进行端到端微调。据我们所知，这是第一篇将价格预测作为文本到分布的自然语言处理任务而非点值预测任务进行探索的论文，我们的方法使得 LLM 的表示层在捕捉与分布相关的特征时得以调整。

使用 LLM 嵌入的回归方法： 一种普遍的方法是使用预训练的 LLM 来生成下游回归任务的文本嵌入 (Imperial, 2021; Gu et al., 2024) 。Tang et al. (2024) 提供的证据表明，即使输入维度增加，LLM 嵌入仍能保持强大的回归性能，而传统的特征工程方法通常会失败。

用于回归的上下文学习： 最新研究揭示了像 GPT-4 和 Claude 这样的大型语言模型通过上下文学习执行回归的惊人能力 (Garg et al., 2022; Vacareanu et al., 2024a) 。他们的工作表明，随着提供的上下文示例数量的增加，回归精度通常会提高。在房地产领域，Chen and Si (2024) 证实了这种行为在价格预测任务中的表现。Lukasik et al. (2024) 进一步通过结合大型语言模型的回归感知推断（RAIL）推进了这一方向，通过优化解码策略增强了零样本数字预测。他们的方法表明，精细校准采样参数可以显著改善回归性能，而无需对模型进行微调。

微调 LLMs 用于回归： 最近的研究表明，微调大型语言模型（LLM）在不同领域的回归任务中效果显著。Morgan and Jacobs (2024) 微调了一种基于 LLaMA 的模型，在化学性质预测中达到了与专门领域模型相当的性能。Zhang et al. (2024) 显示经过微调的基于 BERT 的模型可以有效地从非结构化的房产描述中预测房价，超越仅依赖结构化特征的基准方法。Song et al. (2024) 提出了一种框架，将各种输入格式转换为文本，并将大型语言模型微调为通用的端到端回归器，展示了强大的跨领域性能。

尽管这些研究展示了 LLM 在回归方面的能力，但文献中存在两个空白。最重要的是，现有的工作很大程度上忽略了可以利用 LLM 对文本数据丰富理解的概率回归技术。此外，之前的工作缺乏在相似条件下微调、基于嵌入的方法和上下文学习之间的统一比较。我们主要通过引入一种新的分位数回归方法来解决概率性的空白，该方法使语言模型能够进行不确定性感知的预测，同时提供不同基于 LLM 的回归方法的比较分析。

我们对来自公开可用于研究的不同领域的三个价格预测数据集进行了实验：Amazon 产品、Craigslist 二手车列表和欧洲船艇销售。表格中提供了示例。初步的人工检查发现了许多价格错误的实例，即相对于所售物品而言价格过高或过低（示例在附录中提供）。为了解决这个问题，我们采用了 Claude-3.5-Sonnet LLM，以零

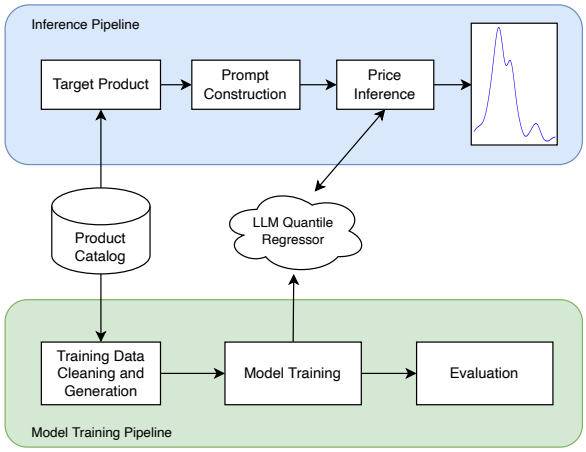

Figure 1: 我们提出的框架的端到端训练和推理流程。

样本的方式识别并移除所有数据集分割中价格不正确的行。关于这一过程的详细信息以及我们的人类评估（94% 的数据表明人类和 LLM 判断一致）确认它既没有移除困难的实例也没有引入偏见，可以在图和附录中找到。本表展示了数据集的分布和每个数据集中被移除的样本数量。

3 基于 LLM 的分位数回归

3.1 问题陈述

给定一个随机变量 X，其实现为 $X = \mathrm{x} \in \mathcal{X}$，代表非结构化的文本输入（例如产品标题或描述）以及其他结构化属性，我们的目标是预测条件分布 $F_{Y|X}(\cdot|\mathrm{x})$，其中 $y \in \mathbb{R}$ 是一个数值结果，比如产品的价格。更正式地说，我们旨在学习一个函数 $f(\cdot;\Theta) : \mathcal{X} \to \mathcal{F}$，该函数将输入映射到条件分布，其中 $\mathcal{F}$ 是 $\mathbb{R}$ 上累积分布函数的空间。这里，Θ 是一个多维参数，例如 LLM 的权重。我们通过一个条件分位数的向量 $\mathrm{q}_{\boldsymbol{\tau}}(\mathrm{x}) = (\hat{q}_{\tau_1}(\mathrm{x}), \ldots, \hat{q}_{\tau_K}(\mathrm{x}))$ 来表示这些分布，针对 K 的预先指定的分位数 $\boldsymbol{\tau} = (\tau_1, ..., \tau_K) \in (0,1)^K$。通过最小化来学习最优参数 Θ^*：

$$\Theta^* = \arg\min_{\Theta} \mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}}[\mathcal{L}(f(\mathrm{x};\Theta), y)]$$

其中 $\mathcal{D}$ 是底层数据分布，而 $\mathcal{L}$ 是概率预测的适当损失函数。图 1 展示了我们提出的使用基于 LLM 的分位数回归进行分布式价格预测的方法的端到端训练和推断流程。

3.2 分位回归头 & 针损失

我们提出在解码器和编码器专用的语言模型架构中添加一个分位数回归头。令 $f(\cdot;\Theta)$ 为以 Θ 参数化的仅解码器的大型语言模型。它接受输入文本（x）的令牌化版本，序列长度为 T，记为 $\tilde{\mathrm{x}} = (x_1, \ldots, x_T)$，并生成隐藏状态 $H \in \mathbb{R}^{T\times D}$，其中 D 为隐藏状态维度。然后我们提取最终隐藏状态 $\mathrm{h}_T = H[T,:]$ 作为序列的总结。对于编码器模型，h_T 是 [CLS] 标记的表示。然后我们用分位数回归头 $g(\cdot;\phi)$ 替换语言模型头，以在给定 $\boldsymbol{h}_T$ 的情况下预测 K 分位数 $\hat{\mathrm{q}} = (\hat{q}_{\tau_1}, \ldots, \hat{q}_{\tau_K})$，并在 appendix C.1 中进行额外的调整。因此，我们的模型结构是：

$$\tilde{\mathrm{x}} = \text{tokenized } \mathrm{x}\,, \tag{1}$$

$$H = f(\tilde{\mathrm{x}};\Theta) \in \mathbb{R}^{T\times D}\,, \tag{2}$$

$$\mathrm{h}_T = H[T,:]\,, \tag{3}$$

$$\hat{\mathrm{q}} = g\big(\mathrm{h}_T;\phi\big) \in \mathbb{R}^K\,. \tag{4}$$

。

斜面损失（也称为分位数损失）允许对预测不足和过度预测进行非对称惩罚。对于在 (4) 中获得的预测分位数 $\hat{\mathrm{q}}$，给定分位数水平 $\boldsymbol{\tau}$ 和真实值 y，我们实现的斜面损失为 $\mathcal{L}_{\boldsymbol{\tau}}(\hat{\mathrm{q}}, y) = (1/K)\sum_{k=1}^{K}\mathcal{L}_{\tau_k}(\hat{q}_{\tau_k}, y)$，其中，

$$\mathcal{L}_\tau(\hat{q}_\tau, y) = \tau(y - \hat{q}_\tau) + \mathrm{ReLU}(\hat{\mathrm{q}}_\tau - \mathrm{y})\,. \tag{5}$$

为了提高优化的稳定性，我们采用了一个平滑的变体：

$$\mathcal{L}^{\alpha}_\tau(\hat{q}_\tau, y) = \tau(y - \hat{q}_\tau) + \alpha \cdot \mathrm{SoftPlus}_\alpha(\hat{\mathrm{q}}_\tau - \mathrm{y}),$$

，其中 $\mathrm{SoftPlus}_\alpha(\mathrm{x}) = \alpha \log(1 + \mathrm{e}^{\mathrm{x}/\alpha})$ 提供了对 ReLU 的可微分近似，具体为 $\alpha \to 0^+$。

4 实验设置

对于所有预测分布的模型，我们取 $K = 200$ 并将区间 $(0,1)$ 分成 K 个等长子区间以获得 $\boldsymbol{\tau}$。我们在附录 D 中讨论了变化的 K 和平滑参数 α 的影响。我们使用生成分布的模型，既用于生成概率输出，也用于点预测。在后者情况下，我们将 $\tau = 0.5$ 的预测分位数作为点估计。此外，作为基准，我们包括仅使用传统平方误差损失进行训练的模型，并使用它们的直接预测进行比较。

使用 LLM 嵌入的基线：文本特征（标题、描述、属性）与适当的字段标记拼接，并使用 Qwen2-7B-instruct 嵌入模型 (Chu et al., 2024) 转换为嵌入。这些嵌入的所有基线模型都以“Qwen-7B-Emb”前缀表示。这些嵌入作为五种模型的输入特征：用于点估计的岭回归和 XGBoost，用于分布预测的分位数回归（具有两个隐藏层），基于对数转换目标进行训练 [*]，

[*]在所有三个数据集中，由于目标是价格，为了在训练期间处理数据集中广泛的取值范围，我们使用了价格的对数变换。

以及两种基于最近邻的分布预测方法。第一种最近邻模型（kNN）使用训练集中的选定邻居的目标值的经验分布来预测分布，而第二种变体采用基于半径的选择标准（rkNN），并设有最低邻居要求。rkNN 的基本原理是确保相比于 kNN 更准确的邻居之间的经验分布。所有超参数通过 5 折交叉验证选择。

微调语言模型具有分位数头：我们对 Mistral-7B (Mistral, 2023)、Phi-3B (Abdin et al., 2024)、Qwen-500M (Bai et al., 2023a) 和 XLM-RoBERTa (Conneau et al., 2019) 模型进行微调，使用第 3 节中描述的分位数回归头。

上下文学习：我们评估了两种最先进的 LLM，Claude-3.5-Sonnet 和 Nova Pro (Anthropic, 2024; Amazon, 2024)，包括零样本和少样本情况下。对于少样本学习，我们实施了三种示例选择策略：(i) 随机采样；(ii) 基于类别的分层采样和 (iii) 基于 Qwen2-7B 嵌入的余弦相似性的相似项采样。后两者利用了领域相似性以期望更好的价格估计（提示如图 7）。

评估指标： 我们使用两组指标。第一组评估点价格估计，包括：(i) 平均绝对百分比误差 (MAPE)，(ii) 加权绝对百分比误差 (WAPE)，其中权重为 1，(iii) 平均百分比误差 (MPE)。

第二组度量衡量的是预测分位数 $\hat{\mathrm{q}}_{\boldsymbol{\tau}}(\mathrm{x}_i) = (\hat{q}_{\tau_1}(\mathrm{x}_i) \leq \cdots \leq \hat{q}_{\tau_K}(\mathrm{x}_i))$ 对于每个输入 x_i 的分布质量。这些度量包括：

(i) 校准误差（CE）衡量预测分位数与它们的理论覆盖度匹配得有多好。$\mathrm{CE} = (1/K)\sum_{k=1}^{K} |\widehat{\mathrm{coverage}}(\tau_k) - \tau_k|$，其中 $\widehat{\mathrm{coverage}}(\tau_k)$ 是测试集中真实值的经验比例，低于 τ_k 分位数。

（ii）连续分级概率技巧评分（CRPSS）衡量预测和真实累积分布函数之间的积分平方差。

（iii）相对置信区间宽度（RCIW）测量预测区间相对于真实值的平均宽度或紧密程度。

对于每个指标，我们报告使用自举重采样（1000 次迭代）得到的 95 % 置信区间：$\mathrm{CI}_{95\%}(M) = [\hat{M}_{(0.025)}, \hat{M}_{(0.975)}]$，其中 $\hat{M}_{(q)}$ 表示指标 M 的自举分布的第 q 个分位数。关于所有指标和训练过程的详细信息见附录 D。

5 结果

5.1 点回归结果

Table 1 列出了我们的主要点回归结果，比较了三个数据集中的所有模型。

经过精调的 Mistral-7B-Quantile 模型在所有数据集上显著优于其他方法。对于 Amazon Products 数据集，Mistral-7B 实现了 16.86 % 的 MAPE，远低于最佳传统基线（Qwen7B-Emb+RkNN-Q）的 42.68 %。这种模式在二手车数据集中尤为明显，其中 Mistral-7B 的 MAPE 为 6.3 %，相较于传统方法高达 235 % 的 MAPE 实现了数量级的改进。这突出了丰富文本理解对于价格回归的重要性。MPE 结果表明，传统方法往往系统性地低估价格，跨数据集的负偏差范围为-24 % 至-135 %。相比之下，Mistral-7B 显示出极小的系统偏差，其 MPE 值接近零：Amazon Products 为-0.88 %，二手车为 0.185 %。

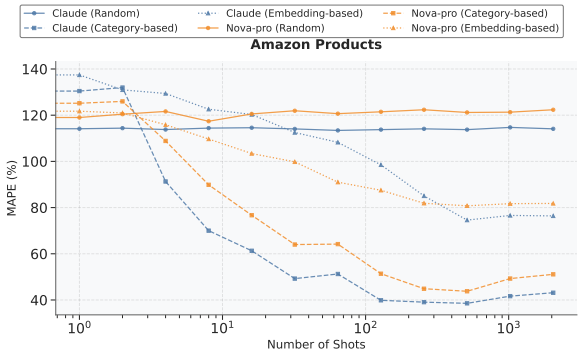

Figure 2: Claude-3.5-Sonnet 和 Nova-Pro LLM 在亚马逊产品数据上的小样本学习性能。

将最佳模型（Mistral-7B-Quantile）与具有点回归头的版本（Mistral-7B-Point）进行比较显示，所有指标都有显著改善，证实了分位数回归分布的中位数比逐点回归是更好的估计。

比较仅编码器（XLM-RoBERTa）和仅解码器（Mistral-7B、Phi-3B、Qwen-500M）架构提供了细致的见解。尽管最佳表现的模型整体上是较大的仅解码器 Mistral-7B，在所有数据集上均实现显著降低的 MAPE（亚马逊产品为 16.9 %，二手车为 6.3 %，船为 21.2 %），但这种优势在相同大小的比较中并不一致。具体来说，基于编码器的 XLM-R-Large 在所有数据集中始终优于相似大小的仅解码器 Qwen-500M，甚至在两个数据集（二手车和船）上超越了更大的 Phi-3B。此外，较小的编码器基础模型 XLM-R-Base 与 Qwen-500M 保持竞争力。在船数据集上 Mistral-7B 和 XLM-R-Large 之间缩小的性能差距（21.2 % 对比 22.7 %）表明，模型性能受到了架构之外多个因素的组合影响，如模型大小、数据和训练方法。鉴于我们的主要目标是识别最强的整体文本到分布预测模型，我们将推迟对建筑对点预测准确性的具体影响进行更详细的研究以量化和隔离这些影响的未来研究。

少样本学习表现不佳。 Figure 2 显示了我们三个数据集上少样本学习方法的结果。即便采用表现最佳的基于类别的采样策略和最佳样本数，Claude 和 Nova-pro 在 Amazon Products

Dataset	Model	MAPE (%) ↓		MPE (%) ↓		WAPE (%) ↓	
		Value	95 % CI	Value	95 % CI	Value	95 % CI
Amazon Products	Mistral-7B-Point	20.81	[20.13, 21.22]	-3.40	[-3.53, -3.30]	22.81	[22.45, 24.67]
	Mistral-7B-Quantile	16.86	[16.15, 17.71]	-0.88	[-1.23, -0.55]	18.32	[17.83, 18.83]
	XLM-R Base-Quantile	41.99	[40.34, 43.86]	-21.73	[-23.72, -20.00]	40.27	[39.09, 41.45]
	XLM-R Large-Quantile	36.52	[34.64, 38.49]	-15.18	[-17.22, -13.22]	37.51	[36.20, 38.74]
	Qwen-500M-Quantile	39.19	[38.01, 40.36]	-6.33	[-7.73, -5.07]	43.15	[42.07, 44.10]
	Phi-3B-Quantile	34.17	[33.14, 35.27]	-5.41	[-6.65, -4.17]	38.17	[37.11, 39.29]
	Qwen-7B-Emb+Ridge	58.97	[57.78, 60.26]	30.36	[32.02, -29.02]	52.72	[51.93, 53.41]
	Qwen-7B-Emb+XGBoost	63.16	[62.22, 64.30]	-32.57	[-33.98, -31.50]	58.01	[57.27, 58.88]
	Qwen-7B-Emb+Quantile	77.97	[76.89, 79.10]	-27.44	[-28.99, -25.88]	76.3	[75.69, 76.99]
	Qwen-7B-Emb+kNN-Quantile	46.86	[45.88, 47.83]	-10.53	[-11.68, -9.27]	54.05	[52.95, 55.04]
	Qwen-7B-Emb+RkNN-Quantile	42.68	[41.66, 43.90]	-10.33	[-11.65, -9.06]	48.03	[46.88, 49.21]
	Claude-3.5-Sonnet (512 category-based shots)	38.50	[36.70, 39.10]	14.32	[14.29, 14.41]	41.40	[40.20, 42.16]
	Nova-Pro (512 category-based shots)	43.77	[40.78, 45.01]	19.12	[18.79, 19.81]	48.13	[46.21, 49.33]
Used Cars	Mistral-7B-Point	9.76	[9.25, 10.67]	-5.40	[-5.89, -4.01]	12.79	[12.65, 13.32]
	Mistral-7B-Quantile	6.30	[6.06, 6.95]	0.19	[0.05, 0.31]	5.40	[5.29, 5.51]
	XLM-R Base-Quantile	11.45	[10.68, 12.44]	-5.71	[-6.62, -4.87]	8.89	[8.62, 9.23]
	XLM-R Large-Quantile	12.84	[12.41, 13.37]	-9.70	[-10.23, -9.24]	10.46	[10.22, 10.74]
	Qwen-500M-Quantile	23.49	[20.61, 26.71]	-4.56	[-8.03, -1.48]	15.93	[15.48, 16.40]
	Phi-3B-Quantile	52.79	[51.83, 53.89]	47.09	[45.90, 48.14]	74.91	[74.50, 75.32]
	Qwen-7B-Emb+Ridge	40.46	[37.95, 43.42]	-18.04	[-21.12, -15.44]	23.04	[22.49, 23.41]
	Qwen-7B-Emb+XGBoost	39.70	[38.10, 41.75]	-16.09	[17.96, -14.41]	26.13	[25.80, 26.60]
	Qwen-7B-Emb+Quantile	235.92	[221.57, 249.55]	-192.67	[-206.41, -178.24]	58.41	[57.85, 58.85]
	Qwen-7B-Emb+kNN-Quantile	79.72	[73.18, 86.87]	-59.39	[-66.09, -52.59]	26.81	[26.33, 27.29]
	Qwen-7B-Emb+RkNN-Quantile	58.18	[52.46, 64.02]	-40.92	[-46.86, -35.39]	21.37	[20.93, 21.84]
	Claude-3.5-Sonnet (2048 random shots)	275.00	[269.12, 280.09]	189.19	[175.21, 195.62]	53.34	[50.78, 56.09]
	Nova-Pro (1024 random shots)	219.67	[167.42, 231.91]	173.07	[156.12, 189.07]	46.44	[42.13, 48.71]
Boats	Mistral-7B-Point	24.01	[23.82, 24.29]	4.10	[2.30, 7.45]	25.82	[24.20, 27.39]
	Mistral-7B-Quantile	21.20	[20.50, 23.39]	2.19	[1.59, 6.65]	23.96	[20.68, 27.69]
	XLM-R Base-Quantile	22.17	[20.26, 24.47]	0.58	[-2.69, 3.52]	23.59	[20.12, 26.78]
	XLM-R Large-Quantile	22.67	[20.85, 24.55]	-4.51	[-7.43, -1.78]	31.05	[24.31, 37.99]
	Qwen-500M-Quantile	62.27	[56.49, 69.12]	16.98	[8.6, 24.73]	77.23	[73.20, 80.75]
	Phi-3B-Quantile	73.83	[71.45, 76.02]	72.89	[70.32, 75.31]	93.64	[92.41, 94.64]
	Qwen-7B-Emb+Ridge	30.77	[28.06, 33.89]	-7.52	[-11.85, -3.81]	28.77	[24.57, 33.38]
	Qwen-7B-Emb+XGBoost	44.56	[40.12, 49.19]	-12.84	[-17.86, -6.93]	42.35	[38.02, 46.85]
	Qwen-7B-Emb+Quantile	131.03	[110.78, 158.77]	-67.61	[-99.21, -44.88]	82.21	[78.75, 84.98]
	Qwen-7B-Emb+kNN-Quantile	77.68	[67.39, 88.29]	-32.36	[-43.49, -20.56]	63.39	[58.06, 67.78]
	Qwen-7B-Emb+RkNN-Quantile	70.96	[61.86, 80.72]	-28.67	[-39.61, -18.10]	56.80	[51.81, 61.44]
	Claude-3.5-Sonnet (2048 random shots)	30.00	[28.97, 31.28]	17.32	[15.16, 19.23]	29.36	[26.16, 30.09]
	Nova-Pro (2048 random shots)	61.01	[55.54, 64.76]	23.22	[21.16, 25.91]	48.79	[45.03, 50.71]

Bold values indicate best performance for each metric and dataset. The ↓ indicates that lower metric values are better.

Table 1: 模型点估计性能比较，使用中位数作为分位数回归模型的点估计。对于小样本 Claude-3.5 和 Nova-Pro LLMs，我们只显示最佳小样本例选择策略以及提供最佳结果的相应小样本数量。

和 Used Boats 上仍然比微调的 Mistral-7B 落后超过 15 %，而在 Used Cars 上则落后超过 200 %。我们的实验还表明，增加示例数在某个点之后，会开始降低模型性能。这一显著的性能差距表明，对于精确的价格预测，微调比精心设计的少样本方法能够产生显著更好的结果。似乎丰富文本数据和价格之间的复杂关系需要比上下文学习所能实现的更彻底的模型适应。

5.2 分布回归结果

Table 2 列出了我们的主要分布回归结果，比较了各种解码器和编码器模型，以及嵌入基线。

如前所述，经过微调的 Mistral-7B-Quantile 模型在所有分布度量和数据集上表现一直很强。它在所有三个数据集上都取得了介于 0.73 到 0.92 之间的最佳 CRPSS 分数，表明相对于参考数据，该模型的概率预测质量很高。Mistral-7B 在亚马逊产品和二手车数据上也取得了最佳的 RCIW 分数，并在船只数据上取得了具有竞争力的分数，显示出分布集中的特点和总体上精确的预测置信区间。总体而言，与较小的模型如 Phi、Qwen-500M 或 XLM-RoBERTa 相比，Mistral-7B 更加一致，能够更好地在不同度量项目中保持平衡。Mistral-7B 和基于 Qwen-7B 嵌入的变体在所有数据集上均表现出非常低的 CE 分数，表明预测的概率

Dataset	Model	CE ↓		CRPSS ↑		RCIW@95 % CI ↓	
		Value	95 % CI	Value	95 % CI	Value	95 % CI
Amazon Products	Mistral-7B-Quantile	0.042	[0.039, 0.044]	0.75	[0.74, 0.76]	0.92	[0.92, 0.93]
	XLM-R Base	0.060	[0.057, 0.064]	0.49	[0.48, 0.51]	2.03	[1.99, 2.09]
	XLM-R Large	0.040	[0.037, 0.043]	0.53	[0.51, 0.55]	1.52	[1.48, 1.57]
	Qwen-500M-Quantile	0.055	[0.051, 0.059]	0.47	[0.46, 0.49]	2.89	[2.83, 3.00]
	Phi-3B-Quantile	0.041	[0.036, 0.046]	0.53	[0.52, 0.54]	2.33	[2.27, 2.38]
	Qwen-7B-Emb+Quantile	0.045	[0.042, 0.048]	0.03	[0.01, 0.04]	16.76	[16.59, 16.92]
	Qwen-7B-Emb+kNN-Quantile	0.01	[0.006, 0.013]	0.34	[0.31, 0.37]	6.14	[6.05, 6.26]
	Qwen-7B-Emb+RkNN-Quantile	0.01	[0.007, 0.012]	0.42	[0.39, 0.44]	6.12	[6.00, 6.29]
Used Cars	Mistral-7B-Quantile	0.054	[0.051, 0.055]	0.92	[0.91, 0.92]	0.20	[0.20, 0.21]
	XLM-R Base	0.157	[0.155, 0.159]	0.80	[0.79, 0.81]	1.02	[1.01, 1.03]
	XLM-R Large	0.185	[0.183, 0.187]	0.80	[0.79, 0.81]	1.01	[1.00, 1.01]
	Qwen-500M-Quantile	0.160	[0.158, 0.162]	0.66	[0.65, 0.66]	4.70	[3.76, 5.50]
	Phi-3B-Quantile	0.395	[0.393, 0.397]	0.04	[0.03, 0.05]	0.99	[0.96, 1.04]
	Qwen-7B-Emb+Quantile	0.020	[0.018, 0.022]	0.01	[0.01, 0.02]	13.41	[12.84, 14.45]
	Qwen-7B-Emb+kNN-Quantile	0.024	[0.020, 0.028]	0.53	[0.52, 0.53]	3.90	[3.75, 4.09]
	Qwen-7B-Emb+RkNN-Quantile	0.022	[0.019, 0.026]	0.62	[0.63, 0.64]	2.64	[2.54, 2.80]
Boats	Mistral-7B-Quantile	0.076	[0.070, 0.084]	0.73	[0.67, 0.77]	1.28	[1.23, 1.35]
	XLM-R Base	0.047	[0.028, 0.066]	0.73	[0.70, 0.77]	1.54	[1.50, 1.59]
	XLM-R Large	0.042	[0.030, 0.051]	0.59	[0.45, 0.69]	1.68	[1.65, 1.72]
	Qwen-500M-Quantile	0.257	[0.237, 0.275]	0.18	[0.03, 0.44]	1.24	[1.17, 1.34]
	Phi-3B-Quantile	0.453	[0.445, 0.461]	0.21	[0.13, 0.35]	0.68	[0.62, 0.74]
	Qwen-7B-Emb+Quantile	0.034	[0.014, 0.056]	0.20	[0.09, 0.23]	33.37	[30.03, 36.55]
	Qwen-7B-Emb+kNN-Q	0.021	[0.011, 0.036]	0.28	[0.19, 0.38]	11.33	[10.41, 12.25]
	Qwen-7B-Emb+RkNN-Q	0.025	[0.013, 0.042]	0.39	[0.30, 0.46]	7.87	[7.03, 8.61]

Bold values indicate best performance. ↓ indicates that lower metric values are better, and ↑ indicates that higher are better.

Table 2: 模型在不同数据集上的分布预测性能比较。CE 衡量预测分位数与其理论覆盖率的匹配程度，CRPSS 评价相对于基准的概率预测质量，而 RCIW 衡量分布预测区间的锐度。

与其理论覆盖率非常吻合。

大语言模型产生更校准的分布。 Mistral-7B-Quantile 模型展示了强的校准能力（CE 在 0.04-0.07 之间），同时保持了更好的置信区间，表明经过微调的大型语言模型在生成良好校准的概率分布方面具有内在优势。Qwen-7B 嵌入变体也实现了极低的校准误差（在亚马逊产品上 CE 为 0.01，在二手车上 CE 为 0.02），显著优于较小的模型如 XLM-RoBERTa，其在亚马逊产品上的 CE 为 0.04-0.06，在二手车上的 CE 为 0.157-0.185。RCIW 模式也揭示了有趣的权衡。虽然基于嵌入的 Qwen-7B 变体实现了优秀的校准能力，但它们在亚马逊产品上的置信区间更宽，RCIW 在 6.12-16.76 之间。然而，Mistral-7B 模型经过微调以适应分位数回归后，可以实现既精确的预测又良好的校准能力，其最佳 RCIW 评分同时保持高 CRPSS，为此提供了证据。

更大的数据导致更好的分布。 大多数模型在较大的 Amazon Products 和 Used Cars 数据集上实现了更好且更一致的分布度量得分，而在更小的 Boats 数据集上的表现则不如前者。与在 Boats 数据上的 RCIW 为 1.28 相比，Mistral-7B-Quantile 在 Used Cars 数据集上实现了 0.2 的更紧密的置信区间。Boats 数据集上的较宽置信区间和较不稳定的模型表现突显了较小样本量对概率预测的不利影响。

分布的定性分析。 我们在图 3 中展示了由我们微调的 Mistral-7B-Quantile 模型预测的价格概率分布函数，并使用高斯核进行平滑处理。我们展示了来自三个数据集的例子，这些数据集具有不同的 MAPE 值（其他例子见附录的第 E.2 节）。在两个亚马逊产品的示例中，预测的中位数价格与实际的真实价格非常接近，分布的最大模态集中在真实价格附近。对于价格较高的二手车数据集，我们观察到更宽的分布，跨越更大的价格范围，并且在船只数据集中看到最大的分布宽度和价格不确定性，这可能是由于该领域价格波动较大或训练数据集较小造成的。

5.3 讨论

分布回归的理论依据。 虽然 Table 1 显示了分布回归（Mistral-7B-Quantile）在所有点指标上相对于点回归（Mistral-7B-Point）的持续优越表现，我们还从理论上讨论了为什么多分位数 LLM 微调在捕获不确定性方面优于点估计微调。用均方误差（MSE）损失来微调 LLMs 训练它们学习条件平均值，忽略了高阶矩和分

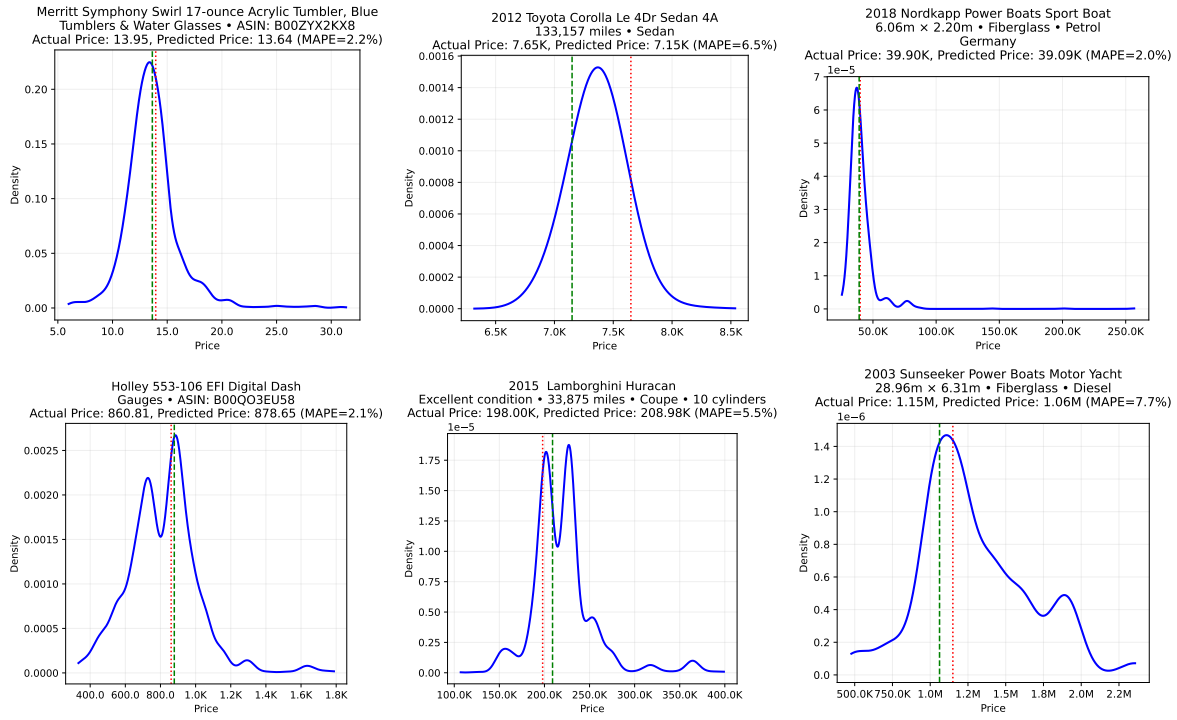

Figure 3: Mistral-7B-Quantile 模型在不同数据集上的预测价格概率密度分布（蓝色 曲线）。每个 x 轴都有不同的比例。 红色 虚线代表真实价格，而 绿色 虚线是预测的中位价格。如图所示，该模型捕捉到包括单峰（顶部行）、双峰（底部行）和右偏（右侧）分布等不同的分布形状。

布形状。这是因为梯度与原始误差 $(\hat{y} - y)$ 成比例，所有修正都将预测推向条件平均值。另一方面，对于分位数 τ 的 Pinball 损失产生该特定分位数 (Koenker and Bassett, 1978) 的一致估计器。对于每个观测，当模型低估 τ -分位数时，梯度以上权重 τ 上移预测；相反，对于高估，梯度以下权重 1- τ 下移。在我们的多分位数方法中，对多个 τ 求和提供了对 (0,1) 中的 τ 上的 pinball 损失积分的离散近似。此积分对应于 CRPS，它是整个分布 (Gneiting and Raftery, 2007) 的严格适当的评分规则，所以最小化它可以恢复真实的条件分布。

从多任务学习的角度来看，我们的方法从跨分位数预测的共享表示中受益；每个分位数水平有效地作为一个相关但不同的预测任务。分位数方法通过两种数学机制产生更好的点估计：首先，$\tau = 0.5$ 分位数损失对离群值具有固有的鲁棒性；其次，在同时训练多个分位数水平时，非中位数分位数损失有效地作为中位数预测任务的正则化项，约束模型在整个分布上表现良好。

按类别的性能分析。 对 Mistral-7B 在亚马逊产品数据中不同产品类别的表现进行分析，揭示了有趣的模式。该模型在价格结构标准化的类别中表现出色，例如在窗膜套件中的 MAPE 低至 4.68 %，该类别的价格跨度也较大，达 $ 200；在钥匙链 & 上达到 5.39 %。虽然在价格范围较窄的类别（例如，机器螺钉：$ 7.35- $ 13.84）或市场细分明确的类别（例如，发动机管理系统）中有高表现力的实例，该模型也能够为价格范围跨度较大的多样化类别做出高质量的价格预测（MAPE 在 6-12 % 之内），例如自定义适配器在价格范围超过 $ 500 时的 MAPE 为 10.43 %，以及车身在价格范围超过 $ 400 时的 MAPE 为 12.11 %。更多详情可参见表格 7 和 8 。

通过训练数据和模型规模提高性能。 Figure 4 展示了随着训练数据增加而获得的明显性能提升。Mistral-7B-Quantile 在 Amazon Products 数据集上显示出显著的扩展优势，MAPE 从 1,000 个样本时的 39.84 % 降低到 100,000 个样本时的 24.3 %。Used Cars 数据集表现出类似的扩展行为，MAPE 从同样范围内的 27.09 % 降低到 19.09 %。Mistral-7B 在所有指标上相较于较小的模型（Phi、Qwen 和 RoBERTa）表现出更优越的性能，显示了模型规模对价格估计精度有显著影响，尤其是在复杂场景中。

训练数据污染。 多个证据表明我们的结果可能不是由于 LLM 预训练和我们的测试数据之间的污染。像 Claude-3.5-Sonnet 这样的最新 LLM 在没有任务特定的微调情况下表现不佳（Table 1 ，图 2 ），这表明在预训练期间对价格

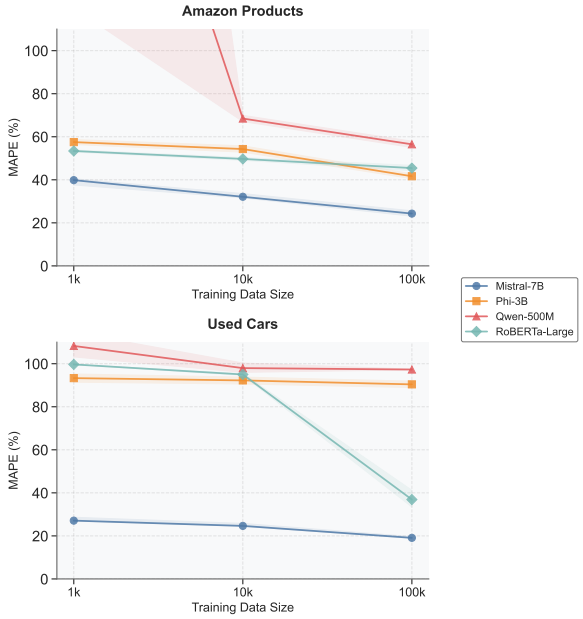

Figure 4: 训练数据规模对模型在两个数据集上的 MAPE 影响（y 轴缩放为 0-100 % 进行比较）。

关系的保留有限。此外，我们的数据缩放实验（图 4）显示随着训练数据的增加，所有数据集的性能都有一致的提升，这说明更多的数据有助于更高程度的学习。然而，我们不能提供确凿的证据证明我们的测试数据在 LLM 预训练期间没有遇到过，尽管有上述证据，污染仍然是一个可能性。

文本到分布建模的实际应用 我们从非结构化文本输入中进行价格分布估计的方法，比传统的点对点回归方法产生了更具信息量的输出。这些价格分布可用于：(i) 捕捉项目间不同程度的价格不确定性，(ii) 提供可解释的概率范围（例如，90 % 置信区间），以及 (iii) 表示如 Figure 3 和 Figure 8 所示的多样分布形态。

6 结论

我们展示了带有分位数回归头的大型语言模型在基于非结构化输入进行概率价格预测方面的有效性，不仅生成校准良好的价格分布，而且相比传统方法还实现了更优的点估计。我们的 Mistral-7B-Quantile 模型在多个数据集上优于传统方法和少量样本的上下文学习，尤其是在模型规模和训练量增加的情况下性能显著提升。我们的研究结果为利用大型语言模型进行概率回归奠定了基础，并展示了它们在使用非结构化输入进行复杂数值预测任务中的能力。

未来有几条有前景的研究途径，例如结合仅解码器模型与传统定价方法的混合架构，明确地融入关于定价和市场动态的领域知识，探索高级 LLM 推理技术，以及构建更具可解释性和可靠性的模型，这些模型可以提供有关它们如何做出定价决策的见解。

基于 LLM 的回归方法也可以应用于许多其他现有的基于文本的任务，例如从新闻文章和社交媒体中进行收益率和波动性的财务预测、情感分析和文本可读性评分。

7 局限性

我们承认本研究存在以下限制。首先，我们没有对参数超过 7B 大小的 LLM 进行微调。其次，尽管我们在这项工作中专注于定价任务，但我们相信我们的分位数回归方法在其他领域中也能够很好地泛化，因为我们的模型结构不包含领域特定的组件。然而，我们没有在其他一般回归领域或非价格预测任务中进行评估。第三，我们确实同意我们的训练数据比较旧，并且我们实验的 LLM 可能在它们的预训练阶段已经见过这些数据。然而，前面讨论的多个实验结果表明这种污染对观察到的结果没有显著贡献。最后，我们的数据集中一些数据可以追溯到 5-10 年前，我们没有详细探讨这些数据对我们上下文学习基准性能的影响。

作者感谢任逸满和 Arman Akbarian 对本项目的初步探索，以及匿名审稿人提供的有益反馈。

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, and Ammar Ahmad Awan et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. Preprint , arXiv:2404.14219.

Amazon. 2024. Amazon nova foundation models. https://aws.amazon.com/bedrock/nova. Accessed February 4, 2025.

Anthropic. 2024. Claude 3.5 sonnet. Accessed February 4, 2025.

Sandeep Arora, James W Taylor, and Ho-Yin Mak. 2023. Probabilistic forecasting of patient waiting times in an emergency department. Manufacturing & Service Operations Management , 25(4):1489–1508.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, and Kai et al. Dang. 2023a. Qwen technical report. arXiv preprint arXiv:2309.16609 .

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. 2023b. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. arXiv preprint arXiv:2306.04637 .

Zsolt Bitvai and Trevor Cohn. 2015. Non-linear text regression with a deep convolutional neural network.

In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) , pages 180–185, Beijing, China. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, and Jared D et al. Kaplan. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems , volume 33, pages 1877–1901. Curran Associates, Inc.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv e-prints , arXiv:2303.12712.

Tingting Chen and Shijing Si. 2024. Predicting rental price of lane houses in Shanghai with machine learning methods and large language models. arXiv preprint arXiv:2405.17505 .

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. arXiv preprint arXiv:2407.10759 .

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2024. A decoder-only foundation model for time-series forecasting. In Forty-first International Conference on Machine Learning .

Neşat Dereli and Murat Saraclar. 2019. Convolutional neural networks for financial text regression. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop , pages 331–337, Florence, Italy. Association for Computational Linguistics.

Ilia D. Dichev, Xinyi Huang, Donald K.K. Lee, and Jianxin Zhao. 2023. Estimating and using distributional forecasts of earnings. Working paper, SSRN. Available at SSRN, Last revised: May 6, 2025.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. In Advances in Neural Information Processing Systems .

Tilmann Gneiting and Adrian E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association , 102(477):359–378.

Nate Gruver, Marc Anton Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. Large language models are zero-shot time series forecasters. In Thirty-seventh Conference on Neural Information Processing Systems .

Wenjun Gu, Yihao Zhong, Shizun Li, Changsong Wei, Liting Dong, Zhuoyue Wang, and Chao Yan. 2024. Predicting stock prices with FinBERT-LSTM: Integrating news sentiment analysis. In 2024 8th International Conference on Cloud and Big Data Computing .

Ragp Gürlek, Francis de Véricourt, and Donald K.K. Lee. 2024. Boosted generalized normal distributions: Integrating machine learning with operations knowledge. Working paper, SSRN. Available at SSRN, Posted: August 1, 2024.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations .

Joseph Marvin Imperial. 2021. BERT embeddings for automatic readability assessment. pages 611–618.

Thomas Kneib, Alexander Silbersdorff, and Benjamin Säfken. 2023. Rage Against the Mean –A Review of Distributional Regression Approaches. Econometrics and Statistics , 26:99–123.

Roger W Koenker and Gilbert Bassett. 1978. Regression quantiles. Econometrica , 46(1):33–50.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. arXiv preprint arXiv:2005.11401 .

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In International Conference on Learning Representations .

Michal Lukasik, Harikrishna Narasimhan, Aditya Krishna Menon, Felix Yu, and Sanjiv Kumar. 2024. Regression aware inference with LLMs. In Findings of the Association for Computational Linguistics: EMNLP 2024 , pages 13667–13678, Miami, Florida, USA. Association for Computational Linguistics.

Mistral. 2023. Mistral 7b. Https://mistral.ai/news/mistral-7b/.

Dane Morgan and Ryan Jacobs. 2024. Regression with Large Language Models for Materials and Molecular Property Prediction (part 1 of 2).

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Proceedings

of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) , pages 188–197.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Anadkat, and Others. 2023. GPT-4 Technical Report. arXiv e-prints , arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155 .

Zihang Qiu, Chaojie Li, Zhongyang Wang, Renyou Xie, Borui Zhang, Huadong Mo, Guo Chen, and Zhaoyang Dong. 2024. EF-LLM: Energy forecasting LLM with AI-assisted automation, enhanced sparse prediction, hallucination detection. arXiv preprint arXiv:2411.00852 .

Xingyou Song, Oscar Li, Chansoo Lee, Bangding Yang, Daiyi Peng, Sagi Perel, and Yutian Chen. 2024. Omnipred: Language models as universal regressors. CoRR , abs/2402.14547.

Eric Tang, Bangding Yang, and Xingyou Song. 2024. Understanding LLM embeddings for regression. CoRR , abs/2411.14708.

Robert Vacareanu, Victor-Andrei Negru, Vlad Suciu, and Mihai Surdeanu. 2024a. From words to numbers: Your large language model is secretly a capable regressor when given in-context examples. arXiv preprint arXiv:2404.07544 .

Robert Vacareanu, Vlad-Andrei Negru, Vasile Suciu, and Mihai Surdeanu. 2024b. From words to numbers: Your large language model is secretly A capable regressor when given in-context examples. CoRR , abs/2404.07544.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. ACM Computing Surveys , 53(3):63.

Hanxiang Zhang, Yansong Li, and Paula Branco. 2024. Describe the house and i will tell you the price: House price prediction with textual description data. Natural Language Engineering , 30(4):661–695.

Yi Zou, Mengying Shi, Zhongjie Chen, Zhu Deng, Zongxiong Lei, Zihan Zeng, Shiming Yang, Hongxiang Tong, Lei Xiao, and Wenwen Zhou. 2025. ESGReveal: An LLM-based approach for extracting structured data from ESG reports. Journal of Cleaner Production , 489:144572.

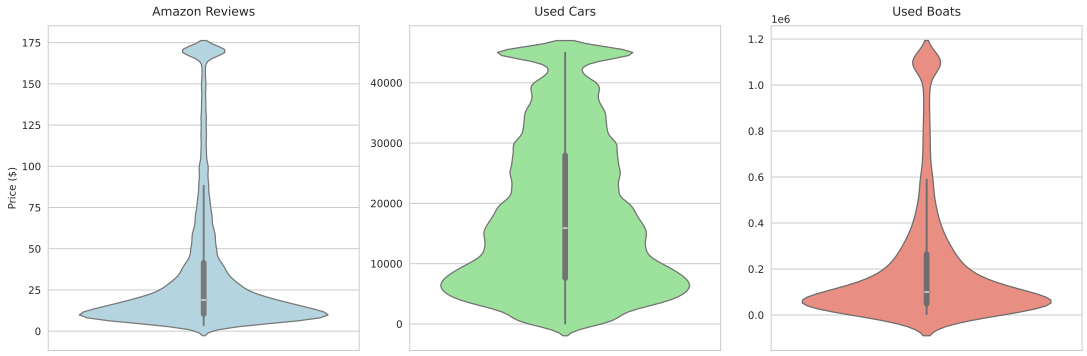

Figure 5: 价格在三种不同数据集中的密度分布：亚马逊产品、二手车和二手船。分布在第 95 百分位被截断以处理异常值。

A

附录

B 数据集详情和示例

我们在 Table 3 展示了我们三个数据集中的输入示例。每个数据条目都包含结构化和非结构化的文本信息。二手船数据集的货币分布如表 4 所示。我们还在图 6 中展示了我们用来清理所有三个数据集并去除包含错误（在第 **??** 节中描述）的行的 LLM 提示。我们在表 6 中展示了此类行的示例。

B.1 验证基于 LLM 的数据过滤

为了解决关于 LLM 过滤是否移除了困难例子或产生意外偏见的潜在问题，我们做出了两个关键观察。首先，我们注意到，执行清理的模型 Claude 在其标记为干净的数据上显示出在零样本和少样本设置中的表现不佳，这提供了它没有选择性地保留易于预测的案例的初步证据。

更为严格地说，我们对一个平衡子集进行的人类评估研究，比较了 LLM 接受和 LLM 拒绝的案例，证实了筛选标准是适当且无偏的。具体而言，我们选择了由 LLM 标记为可接受和不可接受的一个平衡的随机数据子集。独立的人类评估者在不知道 LLM 决策的情况下对这些样本进行了评估。如表 5 所示，人类评估者评估了来自亚马逊数据集的 341 个样本和来自汽车数据集的 153 个样本。结果显示，人类评估与 LLM 判断之间的高度一致性，对于亚马逊和汽车数据集的协议率分别为 95.3 % 和 94.1 %。

为了进一步验证过滤的有效性，我们比较了模型在 LLM 过滤和人工验证子集上的表现。对于 Amazon 数据集，Mistral-7B-Quantile 在 LLM 过滤数据上实现了 16.3 % 的 MAPE（95 % CI：[14.3 % , 18.3 %]），在人为验证数据上则为 43.76 %（95 % CI：[14.9 % , 87.5 %]）。对于汽车数据集，模型在 LLM 过滤和人工验证集上的表现几乎相同，分别实现了 5.82 %（95 % CI：[4.23 % , 7.43 %]）和 5.79 %（95 % CI：[4.22 % , 7.53 %]）。

通过自举抽样比较 LLM 筛选与人工验证测试集的 MAPEs 的 Fisher 检验得到 p 值 0.198 ，表明两个预测集之间的预测准确性没有统计学上的显著差异。该统计证据结合高人类与 LLM 的一致率以及 Claude 在筛选数据上的零样本表现较差，强烈支持我们基于 LLM 的筛选方法的可靠性和无偏性。

B.2 价格分布

图 5 展示了数据集中价格的密度分布。所有分布都显示出显著的右偏模式，虽然集中度和规模有所不同。亚马逊产品的价格在 $ 25 附近出现尖峰，分布相对较窄，表明大部分评论的产品属于可负担的消费品范围。二手车市场则显示出围绕大约 $ 15,000- $ 20,000 的更宽广的分布，价格逐渐向更高价位延展。二手船市场则表现出最大的价格变化，尽管价格达到几百万美元，但核心分布仍集中在较低的价格范围。在此可视化中，所有的分布在 95 百分位处被截断，以修剪离群值。

Dataset Type	Example Data Structure
Amazon Products	{<product> <title>Tubing End Cap Solid Brass Scroll End</title> <description>CAP-off your railing in style with our selection of END CAPS and PLUGS...</description> <brand>Renovator's Supply</brand> <type>Pipe Fittings</type> <attributes>Part Number: 95988, Material: Solid Brass</attributes> </product>, 'price': $34.163}
Used Cars	{<used_car> <model_type>pickup, sierra 1500 crew cab slt, gmc, 2014.0</model_type> <description>Carvana is the safer way to buy a car During these uncertain times, Carvana is dedicated to ensuring safety for all of our customers. In addition to our ...[Removed due to length]</description><size></size><color>white</color> <region>auburn, , al</region><condition>good, clean</condition> <features>cylinders: 8 cylinders, fuel: gas, odometer: 57923.0, transmission: other, VIN: 3GTP1VEC4EG551563, drive: , </features> </used_car>, 'price': $ 33589.548}
Used Boats	{<boat> <boat_type>Flybridge</boat_type> <boat_manufacturer>Galeon power boats</boat_manufacturer> <size>Length: 9.6, Width: 3.0</size> <condition>Used boat, Diesel</condition> <material>GRP</material> <region>Italy ż Lombardia - Trentino Alto Adige ż MARINA DI VERBELLA - LAGO MAGGIORE</region> <year_built>2005</year_built> <price_currency>EUR</price_currency> </boat>, 'price': 68000}

Table 3: 不同数据集的示例数据格式。每个数据集包含非结构化和结构化字段，具有分类和数值属性，涵盖各种商品属性和价格信息。

Currency	Count
EUR	8,430
CHF	980
GBP	298
DKK	180

Table 4: 二手船只的货币分布

Dataset	Total Samples	Both Agree	LLM Acc., Human Rej.	LLM Rej., Human Acc.
Amazon	341	325	14	2
Used Cars	153	144	9	0

Table 5: 人工验证 LLM 清洗后的价格。“Acc.”和“Rej.”代表接受和拒绝。

Table 6: 各数据集被删除的错误价格示例

Dataset	Product Type	Description Summary	Condition	Price
Amazon Products	RAM Memory	16GB (2x8GB) DDR3 RAM for Toshiba Satellite	New	$ 3.28
Amazon Products	Window Insulation Kits	500 sqft (4ft x125ft) of NASA TECH Commercial Grade Reflective Insulation	New	$ 2.85
Used Cars	Mercedes E-Class	2015, 59,749 miles,4MATIC, Blue	Excellent	$ 1.00
Used Cars	Chevrolet Malibu LS Sedan	2015, 79,539 miles, Blue	Clean	$ 165
Boats	Rigiflex Motor Yacht	2017, 4m length, 1.9m width,Switzerland	New	3337 CHF
Boats	Whaly Pontoon boat	2018, 4.35m length, 1.73m width,Italy	New	3300 EUR

C 进一步的建模细节

C.1 确保单调性和连续分位数预测

本节描述了我们可以在分位数回归头中实现的两个结构性添加，之前在 section 3.2 中标记为 $g(\cdot;\phi)$ ，以确保两个性质。首先，是分位数的单调性。具体来说，仅一个分位数回归头和使用钟形损失并不能保证预测分位数 $\hat{q}_{\tau_1}, \hat{q}_{\tau_2}, \ldots, \hat{q}_{\tau_K}$ 会满足 $\tau_i < \tau_j$ 的单调性约束 $\hat{q}_{\tau_i} \leq \hat{q}_{\tau_j}$ 。这可能导致无意义的预测，例如，90 百分位可能低于 80 百分位。第二个问题是分位数分辨率有限。即，在固定的一组 K 分位数水平（例如，$\tau \in \{0.1, 0.2, \ldots, 0.9\}$ ）上进行训练，将预测限制在这些特定水平上，阻碍了对诸如 $\tau = 0.73$ 之类的任意分位数水平的推断。

在下面，我们将描述如何通过增量编码和线性插值的结合来应对这两个挑战。

通过 Delta 编码实现单调性： 与通过 eq. (4) 的回归头直接预测分位数值不同，我们可以调整架构以预测第一个分位数值：$\hat{q}_{\tau_1}$ ，以及连续分位数之间的非负差值：$\Delta_i = \hat{q}_{\tau_{i+1}} - \hat{q}_{\tau_i} \geq 0$ 。

这可以实现为：

$$\mathrm{z}_{\text{deltas}} = [\mathrm{z}_0, \sigma(\mathrm{z}_1), \sigma(\mathrm{z}_2), \ldots, \sigma(\mathrm{z}_{K-1})] \tag{6}$$

$$\hat{\boldsymbol{q}} = \mathrm{CumSum}(\mathrm{z}_{\text{deltas}}), \tag{7}$$

，其中 z 是 h_T ，$\sigma(\cdot)$ 是一个非负激活函数（例如，ReLU 或 SoftPlus ），而 CumSum 表示累积和操作。这样的构建通过设计保证了 $\hat{q}_{\tau_1} \leq \hat{q}_{\tau_2} \leq \cdots \leq \hat{q}_{\tau_K}$ 。

注意，上述修改纯粹是一个体系结构上的修改，通过构建来保证单调性，同时保持损失函数和训练目标完全与 section 3 中描述的相同。网络仍然学习最小化“弹球损失”，只不过它是通过一种使违反单调性成为不可能的结构来完成的。

为了预测训练中未使用的任意水平 $\tau \in (0, 1)$ 的分位数，可以使用相邻训练分位数之间的线性插值。对于查询分位数 τ ，可以找到相邻的训练分位数索引：$i = \lfloor \tau \cdot (K-1) \rfloor$ 和 $i+1$ ，然后计算插值权重：$w = \tau \cdot (K-1) - i$ ，并进行插值：

$$\hat{q}_\tau = (1-w) \cdot \hat{q}_{\tau_i} + w \cdot \hat{q}_{\tau_{i+1}}.$$

这导致在整个范围 $(0, 1)$ 上实现连续的分位数预测，同时保持单调性，因为线性插值能够保持顺序关系。

C.2 小样本学习

小样本学习使模型能够通过有限的训练示例进行预测，这种能力在 LLMs (Wang et al., 2020) 中被证明特别有效。最近的理论工作表明，这种能力，也称为上下文学习，与 Transformer 架构有联系 (Garg et al., 2022; Bai et al., 2023b; Vacareanu et al., 2024a) 。

在我们的定价背景下，少样本学习使大型语言模型（LLMs）能够利用其预训练知识进行价格估算，而只需极少的额外示例。我们通过选择基于类别或各项目制造商与目标产品相似的提示示例来增强这一方法，这类似于检索增强生成（RAG）技术 (Lewis et al., 2021) 。

```
You are an expert in understanding product details and product prices. Given the below information
about a product and its corresponding sale price, judge whether the given price is within a
reasonable range for the given product, or if it is too high or too low.
Also generate a short reason. Your final output should be a single dict within <result> tags with
two keys: price_quality and reason.
[PRODUCT INFO]
Sale Price: [PRICE INFO]
```

Figure 6: Sample LLM prompt that we used to clean up the three of our datasets to remove rows with unreasonably high or unreasonably low prices, with respect to the item contexts.)

```
You are an expert in understanding product details and product prices.
Predict the price in US dollars as a float32 number, for the given set of products. Output a JSON
dict with a key for each input product ID, and a nested dict with a key 'price' containing your
predicted price of the product, and another key 'reason' briefly explaining why your predicted
price is correct.
Put the output JSON dict in <result> tags.

Here are some examples of products and their prices.

[EXAMPLES]
Now predict the price for:
[CONTEXT]
```

Figure 7: Sample prompt for zero shot and few shot LLM based price prediction. This prompt is customized for the Amazon Products dataset, but we used very similar prompts for the other two datasets as well, with minor modifications (e.g., changing references to 'products' to 'used cars' etc.)

我们评估了两种最先进的 LLM，Claude-3.5-Sonnet 和 Nova Pro (Anthropic, 2024; Amazon, 2024) 的零样本和少样本性能。我们实施了三种少样本示例选择策略：（i）随机抽样；（ii）基于类别的分层抽样和（iii）基于 Qwen2-7B 嵌入的余弦相似度的相似项抽样。后两种策略利用领域相似性来可能更好地估算价格。我们根据示例数量 $\{0, 2^0, 2^2, \ldots, 2^{11}\}$，仅受可用数据集大小和 LLM 上下文窗口长度的限制，以分析示例数量和性能之间的关系。所有少样本实验使用一致的提示，如图 7 所示，温度等于 0 。与先前文献 (Vacareanu et al., 2024b) 一致，我们仅使用这些模型进行点估计，因为分布预测需要专门的解码规则 (Lukasik et al., 2024)，这些规则仅限于开源模型。

C.3 使用交叉熵损失

在初步实验中，我们比较了三种微调方法：使用平方误差损失的回归、使用分位数（pinball）损失的回归和使用交叉熵损失的标记预测。回归方法直接优化价格预测，将任务视为一个连续值预测问题，而交叉熵方法将价格视为文本，并遵循传统的下一个标记预测。

在我们的实验中，基于回归的方法明显优于交叉熵方法，平方误差损失在 MAPE 上显示出 1.11 个百分点的改进（95 % 置信区间：[0.40 % , 1.87 %]）。基于这些发现，我们集中于在所有后续实验中进行回归和分位数损失的微调。

D 度量定义和实现细节

我们声明，在本研究中使用的所有公开可用的数据集和模型均符合其许可和使用条款。我们未在其预期用途之外使用任何数据或模型。

对于所有预测分布的模型，我们取 $K = 200$ ，$\boldsymbol{\tau}$ 通过将区间 $(0, 1)$ 划分为 K 等长子区间获得。我们研究了量化数 $K = 10, 50, 200, 500, 1000$ 在三个数据集上的变化影响，发现最初随着 K 增加，性能有所提升，但在某一点后趋于平稳，在我们的情况下，该点是 $K = 200$ 。因此，我们在所有涉及量化回归头的训练模型实验中使用了这个数量。我们使用生成分布的模型来生成概率输出和点预测。在后一种情况下，我们取 $\tau = 0.5$ 处的预测量化值作为点估计。此外，我们包含了仅用传统平方误差损失训练的基线模型，并使用其直接预测进行比较。

我们还调整了平滑参数 α 的值，该参数控制 SoftPlus 函数如何逼近 ReLU 函数。我们试验了从 10^{-5} 到 10^{-1} 的值，但未观察到显著影响。因此，我们选择了 10^{-2} ，以实现接近真实分位数

损失和数值梯度稳定之间的平衡。

我们评估了使用文本嵌入的传统机器学习模型。文本特征（标题、描述、属性）通过适当的字段标记进行连接，并使用通用的 Qwen2-7B-instruct 嵌入模型转换为嵌入。这些嵌入作为五个模型的输入特征：用于点估计的岭回归和 XGBoost，用于分布预测的分位数回归（带有两个隐藏层），在对数变换的目标上进行训练，还有两种基于最近邻的分布预测方法。第一个最近邻模型通过使用训练集中选定邻居的目标值的经验分布来预测分布，而第二种变体则采用基于半径的选择标准，并要求有最小邻居数。所有超参数均通过 5 折交叉验证选择。

我们微调了 Mistral-7B（70 亿参数）、Phi-3B（30 亿参数）和 Qwen-500M（5 亿参数），使用 LoRA（rank=192，alpha=384，dropout=0.1）和 AdamW 优化器（学习率 = $1.0e{-}06$，权重衰减 =0.01），基于第 3 节中描述的分位数头，在对数变换的目标上进行。

我们微调了 XLM-RoBERTa，无论是在基础（279M 参数）还是大型（561M 参数）变体中，并添加了回归头，如第 3 节所述。

我们对 Mistral-7B 进行微调，它是我们集合中最大的 LLM，使用一个回归头以研究分位数预测与点估计的影响。

小样本最新技术大型语言模型： 我们评估了两种最先进的 LLM，Claude-3.5-Sonnet 和 Nova Pro (Anthropic, 2024; Amazon, 2024) 的零样本和少样本性能。对于少样本学习，我们实施了三种示例选择策略：(i) 随机抽样；(ii) 基于类别的分层抽样和 (iii) 基于 Qwen2-7B 嵌入的余弦相似度的类似项目采样。后两者利用领域相似性以实现潜在的更好的价格估计。我们将示例数量作为变量为 $\{0, 2^0, 2^2, \ldots, 2^{11}\}$，仅受限于可用数据集的大小和 LLM 上下文窗口长度，以分析示例数量与性能之间的关系。所有少样本实验使用一致的提示词（附录的图 7）和温度等于 0。与此前文献 (Vacareanu et al., 2024b) 一致，我们仅利用这些模型进行点估计，因为分布预测需要专门的解码规则 (Lukasik et al., 2024)，而这些规则限于开源模型。

对于亚马逊产品和船只数据集，即使采用表现最好的基于类别的采样策略和最佳的采样数量（256），Claude 和 Nova-pro 的 MAPE 也超过了 35 %，远远落后于经过微调的 Mistral-7B 的 MAPE，分别为 16.86 % 和 21 %。在二手车数据集中，表现差距同样显著。对于二手车数据集，虽然 Mistral-7B 达到了 6.3 % 的 MAPE，但少样本方法错误率要高得多：在随机采样时，Claude 和 Nova-pro 的 MAPE 在 230-245 % 之间，而在基于类别的采样中则在 290-305 % 之间。Nova-pro 表现同样糟糕，误差率始终超过 220 %。对于船只数据集，差距有所缩小，但仍然显著。Mistral-7B 的 MAPE 为 21.2 %，仍然大幅优于最佳少样本结果（Claude 在随机采样下的 35 % MAPE）。选择与目标项目在基于 Qwen-7B 嵌入的成对余弦相似基础上相似的少样本示例，也能达到接近随机采样策略 2-3 % 的 MAPE。

我们的实验还揭示了少样本学习性能中的一个有趣模式。与常见的直觉相反，我们的实验还表明，超出某一点增加示例数量会开始降低模型表现。这一发现挑战了更多示例会不可避免地带来更好少样本性能的传统智慧。这种退化可能归因于多个因素，例如模型的上下文窗口大小限制、示例之间的潜在干扰，或从较大示例集中提取相关模式的复杂性增加。这种非单调行为表明，在价格预测任务中必须仔细关注所使用示例的数量和质量，并且存在一个少样本学习的最佳窗口，超过该窗口的额外示例可能会干扰模型有效利用上下文信息的能力。这一观察对少样本学习在定价任务中的实际应用具有重要影响，表明应当仔细关注使用的示例数量，而不是简单地最大化它们。

D.1 评估指标

我们使用两组指标，一组用于评估通过分位数回归模型生成的估计分布，另一组用于点估计。对于每个指标，我们报告带有自助重采样（1000 次迭代）的 95 % 置信区间：$\mathrm{CI}_{95\%}(M) = [\hat{M}_{(0.025)}, \hat{M}_{(0.975)}]$ 其中 $\hat{M}_{(q)}$ 表示指标 M 的自助分布中的第 q 个分位数。

D.1.1 分布质量指标

假设我们有一个大小为 n：$(\mathrm{x}_i, y_i)_{i=1}^n$ 的测试集，对于每个测试点 x_i，我们已经预测了分位数，$\hat{\mathrm{q}}_{\boldsymbol{\tau}}(\mathrm{x}_i) = (\hat{q}_{\tau_1}(\mathrm{x}_i) \leq \cdots \leq \hat{q}_{\tau_K}(\mathrm{x}_i))$。

校准误差 (CE): CE 衡量预测分位数与其理论覆盖率的匹配程度：$\mathrm{CE} = (1/K)\sum_{k=1}^{K} |\widehat{\mathrm{coverage}}(\tau_k) - \tau_k|$。其中，$\widehat{\mathrm{coverage}}(\tau_k)$ 是测试集中真实值低于 τ_k 分位数的经验比例。

Category	MAPE [%]	Size	Price Range [$] [Min, Median, Max]
Camera Lenses	34.75 [20.57, 44.61]	6	[7.78, 36.60, 294.95]
Tools	33.92 [23.22, 44.19]	6	[7.99, 15.93, 84.95]
Bakeware Sets	33.29 [25.07, 40.03]	8	[3.99, 8.09, 130.48]
Compressors	33.09 [25.68, 39.96]	13	[29.99, 165.00, 395.65]
All-Purpose Labels	30.19 [15.89, 43.41]	7	[4.99, 14.95, 33.29]
Platters	29.86 [20.35, 38.93]	6	[14.99, 26.46, 69.99]
Pickups & Pickup Covers	29.86 [22.64, 36.16]	9	[6.04, 12.40, 219.00]
Lighting Assemblies	29.66 [15.11, 44.26]	6	[14.98, 35.67, 173.43]
Hard Hat Accessories	29.40 [17.66, 42.69]	6	[3.99, 4.99, 11.09]
Internal Hard Drives	29.17 [19.69, 38.61]	12	[14.99, 52.50, 599.99]

Table 7: 亚马逊产品数据集中具有高 Mistral-7B MAPE 的类别示例（最小大小 > 5）

连续排名概率技巧评分（CRPSS）： 这个指标是众所周知的 CRPS 的一种无量纲版本，用于测量预测和真实累积分布函数之间的积分平方差异：

$$\mathrm{CRPS} = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\left(\hat{F}_{\mathrm{x}_i}(r) - \mathbf{1}_{y_i \le r}\right)^2 dr\,,$$

其中 $\hat{F}_{\mathrm{x}_i}$ 是使用 $\hat{\mathrm{q}}_{\boldsymbol{\tau}}(\mathrm{x}_i)$ 估计的 CDF。作为一个适当的评分规则，CRPS 当且仅当预测分布与真实分布一致时收敛于零 (Gneiting and Raftery, 2007) 。我们报告（无量纲的）技能得分

$$\mathrm{CRPSS} = 1 - \left(\frac{\mathrm{CRPS}_{\text{model}}}{\mathrm{CRPS}_{\text{reference}}}\right).$$

，其中参考是训练目标的经验分布。

相对置信区间宽度 (RCIW)。 RCIW 衡量相对于真实值的预测区间的平均宽度：

$$\mathrm{RCIW}_{\gamma} = \frac{100}{n}\sum_{i=1}^{n}\frac{U_i^{\gamma} - L_i^{\gamma}}{|y_i|}$$

其中 $[L_i^{\gamma}, U_i^{\gamma}]$ 是 x_i 的预测 $(1-\gamma)$ CI。RCIW 反映了分布的尖锐度，其中较小的值表示区间更紧密。

我们报告: MAPE (平均绝对百分比误差), WAPE (加权绝对百分比误差), 和 MPE (平均百分比误差):

D.2 计算基础设施

我们使用了 AWS EC2 基础设施来运行我们所有的实验。我们估计用于所有模型训练和评估的 GPU 时间约为 2000 小时。我们还使用了人工智能助手来帮助部分代码编写工作。

E 模型性能的详细分析

在本节中，我们对我们的 Mistral-7B-Quantile 模型在不同产品类别中的表现进行了检验，并分析了模型捕获的分布模式。首先，我们展示了按类别划分的预测准确率，揭示了哪些产品类型在价格预测中最具（或最不具）挑战性。然后，我们探讨了模型如何捕捉反映不同产品市场动态的各种分布形状。

E.1 按类别划分的性能分析

我们在表格 7 和 8 中提供了我们最佳模型在每个数据集不同类别上的详细性能分析。

E.2 价格预测中的分布模式

图 3 和图 8 中的概率分布展示了不同的模式，这些模式反映了不同产品类别的市场动态。

Category	MAPE [%]	Size	Price Range [$] [Min, Median, Max]
Window Tinting Kits	4.68 [2.74, 6.75]	19	[24.49, 39.49, 283.94]
Keyrings & Keychains	5.39 [2.59, 8.12]	6	[5.99, 8.09, 10.19]
CV Boots & Joints	5.69 [3.38, 8.31]	11	[11.50, 11.88, 69.99]
Exhaust	5.80 [2.69, 9.22]	8	[15.72, 122.78, 719.48]
Machine Screws	6.01 [2.65, 9.60]	8	[7.35, 9.96, 13.84]
Socket Wrenches	6.30 [2.32, 11.21]	6	[8.51, 14.65, 108.38]
Engine Management Systems	6.65 [3.72, 10.82]	19	[15.22, 69.95, 69.95]
Inkjet Printer Paper	6.79 [3.66, 9.69]	6	[9.50, 29.42, 152.26]
License Plate Frames	6.94 [2.62, 13.28]	11	[5.66, 16.99, 29.99]
Highball Glasses	6.94 [3.08, 11.00]	6	[34.46, 47.27, 110.36]
Touchup Paint	7.37 [5.76, 9.05]	68	[8.25, 15.30, 71.92]
Keychains	9.70 [7.79, 11.83]	71	[5.79, 10.99, 55.99]
Frames	9.91 [8.64, 11.18]	231	[4.99, 14.99, 95.00]
Custom Fit	10.43 [9.19, 11.68]	175	[18.99, 119.00, 599.00]
Body	12.11 [10.63, 13.60]	136	[6.99, 43.49, 409.85]

Table 8: 在亚马逊产品数据集中，Mistral-7B MAPE 低的类别示例（最小规模 > 5）

单峰分布： 产品如 Merritt 杯子、丰田卡罗拉、福特野马和婚礼宾客簿展示了单峰分布。这些标准化产品通常具有已建立的市场价格，且价格波动较小。狭窄且对称的分布表明，由清晰市场细分和标准化特征推动的可预测定价，这体现在模型对这些项目较低的预测误差中（MAPE 介于 1.2 % 至 6.5 %）。

双峰分布： 几种产品显示出双峰现象，包括 Holley EFI 仪表、空调压缩机，以及在不同程度上的兰博基尼 Huracán 和丰田 Tacoma。这种双峰现象可能反映了不同的市场细分。对于汽车零部件（仪表、压缩机），这两个峰可能代表新市场与翻新/二手市场在不同价格点上运作。对于车辆，不同的配置级别、车型年份或状况类别（例如，认证的二手车与标准二手车）会形成独立的价格集群。最后，丰田 Tacoma 的双峰模式可能反映了基础工作卡车与配置齐全的消费车型之间的价格差距。

豪华游艇（Sunseeker 游艇、Storebro 和 Baikai 飞桥）呈现出右偏分布且具有较重的尾部。这种模式与高端市场的特征一致，其中，基本型号形成主要峰值，大量的定制选项、稀有特征或完美/收藏家的条件形成了长尾。此外，极端尾部（特别是在 Baikai 船中明显，价格达到 $ 500K 以上）可能代表高度定制或稀有的配置。

分布形状与预测准确性之间的相关性值得注意。具有单峰分布的标准化产品实现了较低的预测误差，而具有复杂、偏斜分布的奢侈品则表现出更高的不确定性（MAPE 高达 37 %），这可能是由于为高变动性、定制化产品定价本身就很困难。

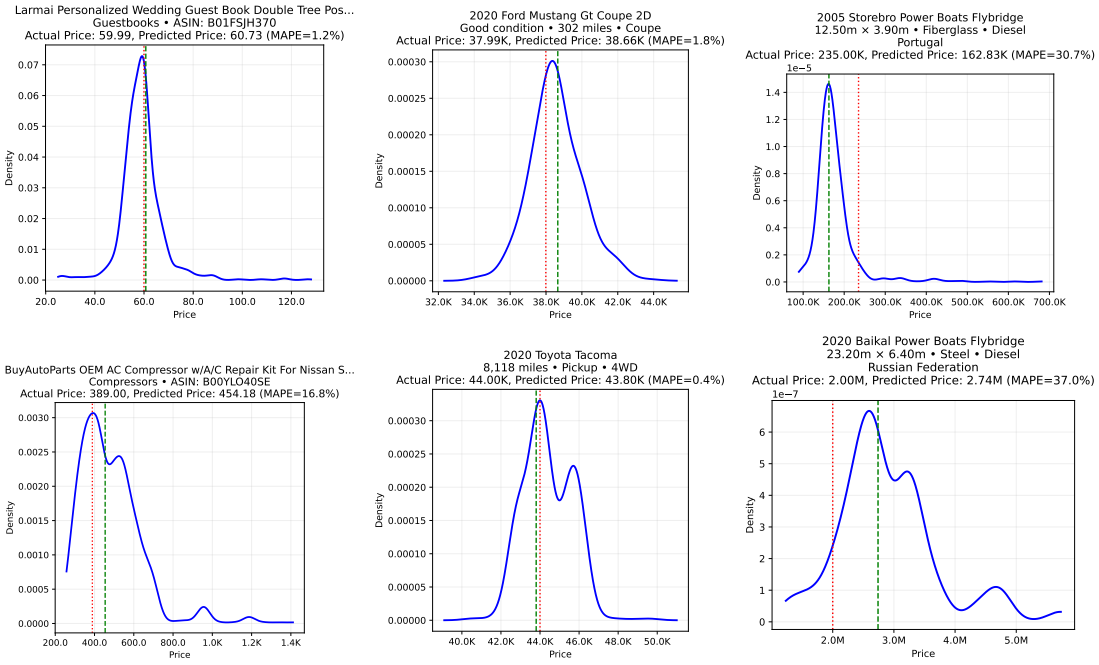

Figure 8: Mistral-7B-Quantile 模型在不同数据集上的价格概率密度分布（蓝色 曲线）。每个 x 轴都有不同的尺度。红色 虚线代表真实的价格，而 绿色 虚线是预测的中位价格。如所示，该模型捕捉到不同的分布形状，包括单峰（顶部行）、双峰（底部行）和右偏（右侧）分布。