LATTE：为银行客户学习对齐的交易和文本嵌入

Egor Fadeev[1], **Dzhambulat Mollaev**[1], **Aleksei Shestov**[1], **Dima Korolev**[1], **Omar Zoloev**[1], **Ivan Kireev**[1], **Andrey Savchenko**[1], **Maksim Makarenko**[1]
[1]Sber AI Lab

Abstract

从客户历史交流的序列中学习客户嵌入是金融应用的核心。虽然大型语言模型（LLMs）提供一般的世界知识，但它们直接用于长事件序列在实际管道中是计算昂贵且不切实际的。在本文中，我们提出了**LATTE**，一种对比学习框架，该框架将原始事件嵌入与冻结的 LLMs 中的语义嵌入对齐。行为特征被总结成简短提示，通过 LLM 进行嵌入，并通过对比损失作为监督使用。与传统的通过 LLM 处理完整序列的方法相比，该方法显著减少了推理成本和输入大小。我们的实验表明，在真实世界的金融数据集中，我们的方法在学习事件序列表示方面优于最新技术，同时仍可部署在对延迟敏感的环境中。

1 介绍

从结构化事件序列中学习表示仍然是医疗保健、教育、电子商务，特别是金融领域工业应用中的核心挑战。在银行业中，数百个时间顺序、有高维度且通常稀疏的表格数据流（如交易记录、支付历史和客户互动）在不断生成。这些数据支撑着广泛的对业务至关重要的任务，包括客户流失预测、风险评估、信用评分和个性化目标营销。尽管有丰富的客户互动数据，但用于这些下游任务的高质量标签仍然有限。这种及时监督的缺乏阻碍了在生产环境中监督学习的可扩展性。这也突显了需要自监督方法直接从原始行为序列中得出稳健且语义丰富的表示。目前用于事件序列建模的自监督方法采用了诸如对比学习、下一个事件预测和潜在序列建模技术（例如，对比预测编码）等目标，旨在捕捉时间依赖性和用户意图，而不依赖人工监督。

近年来大型语言模型（LLMs）的进步为从事件序列中增强表示学习提供了新机会。这些模型在多样且大规模的语料库上训练，能够对通常在结构化事件数据集中隐含或缺失的元素进行编码，例如关于行为模式、时间动态和领域知识的丰富语义先验。利用这些外部知识可以显著提升用户表示的质量，尤其是在金融应用中 [43] 。受这种潜力的启发，最近的研究探索了将 LLMs 适应于结构化数据的方法。例如，TALLRec [19] 、LLM-TRSR [20] 和 HKFR [21] 将 LLMs 丰富的文本理解能力转移到推荐系统中；TabLLM [18] 针对表格分类任务；TEST [17] 和 Time-LLM [16] 处理时间序列；而 ESQA [15] 将 LLMs 应用于事件序列问答。

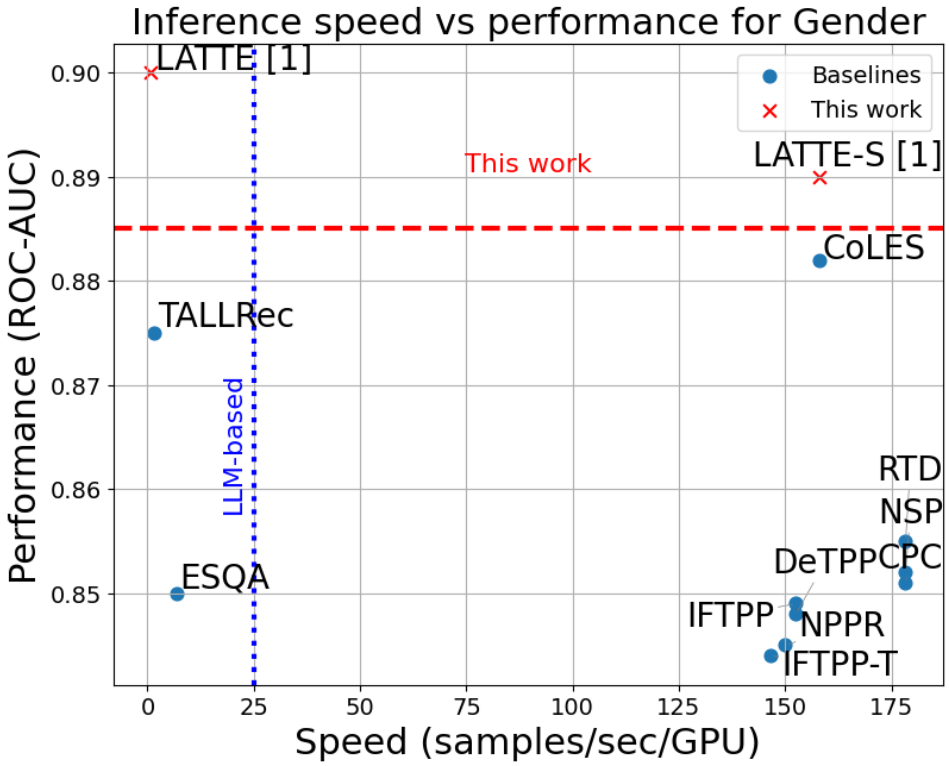

Figure 1: 性别预测任务中 ROC-AUC 性能和推理速度（样本/秒/GPU）的评价指标。比较的方法包括 **LATTE** [1]，**LATTE**-S[1]，CoLES [1] RTD [24]，CPC [23]，NSP [25]，NPPR [22]，DeTPP [33]，IFTPP [34]，IFTPP-T [34]，ESQA [15]，和 TALLRec [19]。

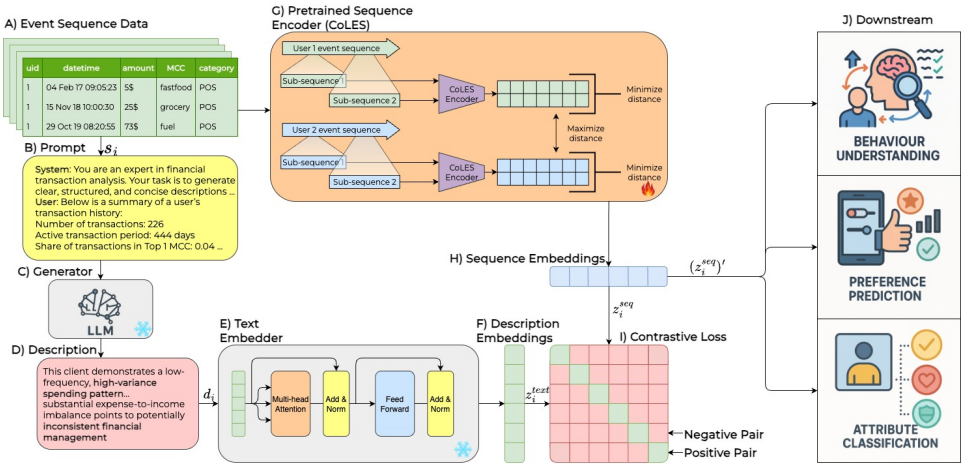

Figure 2: 提出的 **LATTE** 流程概述。(A) 事件序列作为输入数据源。(B) 设计摘要提示以查询事件序列。(C) LLM 生成器根据提示生成自然语言描述。(D) 生成的文本描述捕捉事件序列的显著特征。(E) 文本嵌入器将描述转换为向量表示。(F) 这种描述嵌入编码了生成的文本。(G) 同时，序列编码器嵌入原始事件序列。(H) 生成的序列嵌入捕捉结构和时间模式。(I) 对比对齐模块训练模型以在共享表示空间中对齐文本和序列嵌入。(J) 对齐后的嵌入可用于各种下游任务，如分类、检索或预测。

然而，将大型语言模型（LLM）轻率地直接应用于序列化的顺序表格数据，由于每个用户的大量标记数量，导致了大量的计算开销。例如，典型的银行交易数据集通常每个用户包含数百条记录，每条记录序列化为几十个标记，超出了实际的上下文窗口限制，并增加了推理和训练时间。图 1 显示，基于 LLM 的模型如 TALLRec [19] 和 ESQA [15] 的推理速度不超过每秒 10 个用户，这极大地限制了它们在银行生产环境中的适用性。

现有的方法来缓解这个问题主要分为两类。第一类通过用通用的 LLMs [21, 20] 对用户历史进行总结来减少上下文长度，这存在丢失对准确建模至关重要的领域特定信息的风险。第二类方法试图通过将序列编码为非文本格式 [17, 16] 来绕过上下文长度限制。然而，这些表示通常会丢弃项目描述中存在的语义内容（例如，交易类别或商户详细信息）。此外，随着用户历史变得更长，这些模型要么会导致计算成本增加，要么由于模型容量有限而导致性能下降。

为了解决这些限制，我们提出了 **LATTE**[1]，这是一个用于学习对齐的交易和文本嵌入的可扩展框架。我们不将整个序列输入大语言模型 (LLM)，而是提取简洁的客户级统计数据，并使用指令调优的 LLM 生成自然语言摘要。这些摘要作为弱标签，与经过轻量级编码器的预训练序列嵌入通过对比学习对齐。在推断时，**LATTE** 支持两种模式：一种是独立编码器（**LATTE**-S），保留 LLM 级别的语义而不增加额外开销；另一种是结合编码器（**LATTE** ），融合文本和结构表示。

[1]源码可在 https://anonymous.4open.science/r/LATTE-3BD2/ 下载

为了评估性能和效率之间的折衷，我们引入了一个优值（FOM），该优值比较了性别预测任务的标准银行数据集上的 ROC-AUC 和推理速度 [3]（图 1）。我们评估了 **LATTE** 的两种变体，结合了两种嵌入策略：仅结构的（**LATTE**-S）以及与文本特征拼接的（**LATTE** ）。在性能方面，所有版本的 **LATTE** 均优于典型的基线金融方法。仅结构的变体（**LATTE**-S）在性别任务上实现了 0.899 的 ROC-AUC，推理速度为 162 样本/秒/GPU，超过了如 TALLRec 和 ESQA 这样的基于 LLM 的模型，速度快了 14× 倍。在所有类型的文本编码器中，结合变体（**LATTE** ）始终实现了最高的整体 ROC-AUC。

2 提出的方法

我们旨在通过引入一种从文本行为描述中提取的辅助模态来提高从事务事件序列中学习到的表示的质量。为此，我们提出了一个三阶段的流程 **LATTE** ，如图 2 所示，该流程将原始交易序列映射为适合下游任务的丰富嵌入。

令 $T_i = x_1, x_2, \ldots, x_n$ 表示客户端 i 的交易序列，其中每个 x_j 包含带有时间戳的分类和数值属性（例如，金额，商户类别）。如图 2 A 所示，我们首先计算一个汇总特征向量 s_i ，该向量聚合了序列中的行为模式：活动频率、商户多样性、交易类型、时间覆盖范围和收入-支出结构。分类统计被转化为有意义的文本描述，而不是原始索引。有关提示模板和生成描述的示例，请参见附录 C 。该概要随后被序列化为文本提示（图 2 B），并传递给一个冻结的指令调优 LLM（图 2 C），以生成关于客户端行为的自然语言描述 d_i （图 2 D）。该提示包含一个系统消息和一个设计为引发一致且可解释响应的 s_i 结构化表示。

同时，原始序列 T_i 通过基于 GRU 的序列编码器处理（图 2 G），该编码器在自监督的 CoLES 目标下训练 [1] ，每个训练样例由两个重叠的子序列（正样本）和来自其他客户端的对比负样本组成。这生成了序列嵌入 z_i^{seq} ，优化以捕捉特定客户端的行为动态。与此同时，描述 d_i 通过一个冻结的多语言句子编码器（图 2 F）传递，通过平均池化生成文本嵌入 z_i^{text} （图 2 E）。

然后使用三种跨模态对比损失之一（图 2 I）来对齐嵌入 z_i^{seq} 和 z_i^{text} ，同时保持文本编码器不变。更新后的事务嵌入 $\left(z_i^{\text{seq}}\right)'$ （图 2 H），与文本表示对齐，并评估它们在下游任务上的表现，例如用户流失预测（图 2 J）。

为了实现跨模态对齐，我们引入了三个对比头：对称 Softmax、成对二元和正交正则化。每一个都受到了多模态学习 [6, 7, 32] 的启发。这些头在对齐几何和正则化上有所不同，我们将它们分别称为 **LATTE** [1]、**LATTE** [2] 和 **LATTE** [3]。三个头都使用冻结的文本编码器，仅更新事务编码器。使用所得 $\left(z_i^{\text{seq}}\right)'$ 对下游任务进行评估，而文本模态通过丰富分类属性的上下文知识提供语义上扎根的对齐，而序列编码器单独无法解释这些上下文知识。

LATTE [1]：对称 Softmax 对比头利用对称 InfoNCE 风格损失促进模态之间的双向对齐：

$$\mathcal{L}_{\text{softmax}} = \frac{1}{2}(\mathcal{L}_{\text{seq}\to\text{text}} + \mathcal{L}_{\text{text}\to\text{seq}}), \tag{1}$$

其中

$$\mathcal{L}_{\text{seq}\to\text{text}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\langle z_i^{\text{seq}}, z_i^{\text{text}}\rangle/\tau)}{\sum_{j=1}^{N}\exp(\langle z_i^{\text{seq}}, z_j^{\text{text}}\rangle/\tau)}.$$

第二项 $\mathcal{L}_{\text{text}\to\text{seq}}$ 的定义相似，只是序列和文本的角色互换。这里 $\langle\cdot,\cdot\rangle$ 表示序列（z_i^{seq}）和对应文本（z_i^{text}）的 L2 归一化嵌入之间的余弦相似性，τ 是控制相似性分布锐度的温度超参数。

LATTE [2]：成对二元对比头在批中所有序列描述对之间应用基于 Sigmoid 的监督：

$$\mathcal{L}_{\text{sigm}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\log\left(\frac{1}{1+e^{y_{ij}(-\tau z_i^{\text{seq}}\cdot z_j^{\text{text}}+b)}},\right) \tag{2}$$

其中 $y_{ij}=1$ 对于 $i=j$（正对），否则为 $y_{ij}=-1$（负对），而 b 是可训练的偏差。

LATTE [3]：正交正则化对比头通过一个几何正则项增强了基于 softmax 的损失，该项解耦了特定模态和共享信息。为此，我们引入了一个辅助投影头，将每个事务嵌入 z_i^{seq} 映射到由两部分组成的表示：Z^{shared}，其捕捉与文本信息对齐的成分，以及 Z^{spec}，其保留特定于事务模态的信息：

$\mathcal{L}_{\text{ortho}} = \left\|(Z^{\text{shared}})^{\top}Z^{\text{spec}}\right\|_F^2$ 所在的

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{softmax}} + \lambda_{\text{ortho}}\cdot\mathcal{L}_{\text{ortho}}, \tag{3}$$

。

这种分离通过惩罚共享和特定子空间之间的相关性来促进特征的解耦，其中 λ_{ortho} 控制这种正则化的强度，从而强调保留特定模态的信息。

3 实验装置

本节提供了我们实验设置的核心细节，包括验证策略、数据集和基线方法。附加的实现细节在附录 A 中提供。

3.1 数据

我们在三个真实世界的数据集上评估我们的方法，这些数据集包含来自主要金融机构的匿名信用卡交易序列。每个数据集包括具有数值和类别属性（例如，金额、商家类别、交易类型）的客户级别序列，并包括一个专用于表示学习的未标记子集。Churn [2] 包括约 10K 名由未来不活跃性标记的 Rosbank 客户。Gender [3] 和 Age Group [4] 由 Sberbank 提供，分别包含 15K 和 50K 名用人口统计标签注释的客户。

3.2 验证策略

每个数据集被划分为互不重叠的训练和测试分区，其中保留 10 % 的已标记客户端用于评估。剩下的 90 % 已标记用户，与所有可用的未标记用户一起用于训练嵌入模型。为了评估学习表示的质量，我们采用 5 折交叉验证程序。具体来说，训练数据的标记部分被分为五个大小相等的折叠。对于每个

折叠 v，我们：（1）在剩下的四个折叠的嵌入上训练一个 LightGBM [38] 分类器，以及（2）在保留的测试折叠上评估它，计算一个下游性能指标 M_v。对于二元分类任务（流失，性别），我们报告 ROC-AUC；对于多类年龄预测，我们报告分类准确性。最终性能总结为 $\mu_M \pm \sigma_M$，其中 μ_M 是所有五个折叠的均值，σ_M 是标准差。

3.3 基线

我们将 **LATTE** 与五种不同方法学家族的基线进行比较：
事件序列模型。CPC [23] 通过对比损失学习从过去的上下文中预测未来的表征。CoLES [1] 通过使用 InfoNCE 对齐重叠的子序列来提高时间一致性。NPPR [22] 采用自回归训练，具有双重目标：预测下一个事件和从掩码序列中重构之前的事件。所有模型都在原始事件序列上运行，没有外部模态。
时间点过程模型。DeTPP [33] 使用参数点过程建模事件的时间和类型。IFTPP 和 IFTPP-T [34] 使用 Transformer 和 GRU 框架，将 MAE 和分类损失结合。
自然语言处理目标。RTD [24] 随机替换 15 % 的事件标记并预测原始标记；NSP [25] 将 BERT 的下一句预测扩展到事件序列。
基于 LLM 的方法。我们改编了 TALLRec [19] 和 HKFR [21]，从推荐系统进行用户嵌入提取，通过在用户-项目历史的序列化上微调 LLM 进行下一个标记预测。嵌入是通过对最终层进行均值池化得到的，遵循最近的最佳实践 [46, 42]。此外，我们用 LLaMA 3.2 3B [47] 替换了旧的框架（例如，LLaMA 7B [49]，ChatGLM-6B [48]）以利用架构改进和效率。
表格特征聚合。作为一个非顺序基线，agg 使用汇总统计，如平均值、标准差、最小-最大值和分组频率统计（针对分类特征），对事务特征进行聚合。

4 实验结果

4.1 主要结果

表格 1 展示了客户端嵌入在三个下游任务中的表现：流失预测（AUC）、年龄组分类（准确率）和性别预测（AUC）。虽然像 CPC、RTD 和 NSP 这样的传统自监督目标落后于 CoLES 和 TALLRec 等更强的基线，但我们提出的 LATTE 框架的几个变体在各个任务中表现出明显的提升。**LATTE** [1] 在流失预测中取得了最好的结果（0.868 AUC），而 **LATTE** [3] 在年龄组分类（0.678 准确率）和性别预测（0.902 AUC）上领先。这些改进表明，加入合成文本监督导致了交易行为的表示显著增强。

4.2 运行时分析

在银行应用中，由于需要在有限获取 GPU 的硬件环境下处理数百万个多来源的用户记录，因此推理速度至关重要。在本节中，我们研究推理资源利用（计算时间、内存利用）与所提出方法的性能相比于基线架构的一个重要权衡。

Model	Churn (AUC)	Age Group (Acc)	Gender (AUC)
agg	0.827 ś 0.010	0.629 ś 0.002	0.877 ś 0.004
CPC	0.792 ś 0.015	0.602 ś 0.004	0.851 ś 0.006
RTD	0.771 ś 0.016	0.631 ś 0.006	0.855 ś 0.008
CoLES	0.841 ś 0.005	0.644 ś 0.005	0.882 ś 0.004
NSP	0.828 ś 0.012	0.621 ś 0.005	0.852 ś 0.011
NPPR	0.845 ś 0.003	0.642 ś 0.001	-
DeTPP	0.823 ś 0.002	0.632 ś 0.004	-
IFTPP	0.828 ś 0.004	0.632 ś 0.003	0.863 ś 0.003
IFTPP-T	0.814 ś 0.004	0.620 ś 0.002	0.852 ś 0.005
TALLRec	0.839 ± 0.003	0.659 ś 0.004	0.875 ± 0.004
HKFR	0.823 ± 0.006	-	-
LATTE [1]	0.867 ś 0.004	0.666 ś 0.005	0.900 ś 0.005
LATTE [2]	0.868 ś 0.006	0.671 ś 0.005	0.901 ś 0.005
LATTE [3]	0.857 ś 0.003	0.678 ś 0.003	0.902 ś 0.006

Table 1: 客户端嵌入在下游任务中的表现。加粗表示最佳结果；underline 表示次佳结果；double underline 表示第三佳结果

图 3 展示了在 Churn 数据集上的表现。生成模型（**LATTE** [1 , 2 , 3], TALLRec）实现了最高的 ROC-AUC 分数，超过 0.868，但代价是极低的推理速度——每秒仅处理少量样本——以及超过 30 亿参数的高内存消耗。相比之下，轻量级的 **LATTE**-S 变体在每秒处理超过 160 个样本的小体积（仅 440 万参数）下，获得略低的 ROC-AUC 值。这些模型匹配了对比基准的效率，提供了准确性、速度和内存占用之间的良好平衡。正如附录 B （图 4 ）所示，在年龄预测任务上也出现了类似的趋势。

4.3 消融研究

Method	Churn (AUC)	Age Group (Acc)	Gender (AUC)
Descriptions	0.718 ś 0.004	0.552 ś 0.002	0.692 ś 0.006
CoLES + descriptions	0.853 ś 0.009	0.662 ś 0.005	0.895 ś 0.008
LATTE-S[1]	0.852 ś 0.005	0.657 ś 0.007	0.890 ś 0.008
LATTE [1]	0.867 ś 0.004	0.666 ś 0.005	0.900 ś 0.005
LATTE-S[2]	0.858 ś 0.007	0.662 ś 0.006	0.894 ś 0.009
LATTE [2]	0.868 ś 0.006	0.671 ś 0.005	0.901 ś 0.005
LATTE-S[3]	0.844 ś 0.005	0.663 ś 0.004	0.899 ś 0.001
LATTE [3]	0.857 ś 0.003	0.678 ś 0.003	0.902 ś 0.006

Table 2: 消融：对比微调和模态拼接对下游任务的影响。

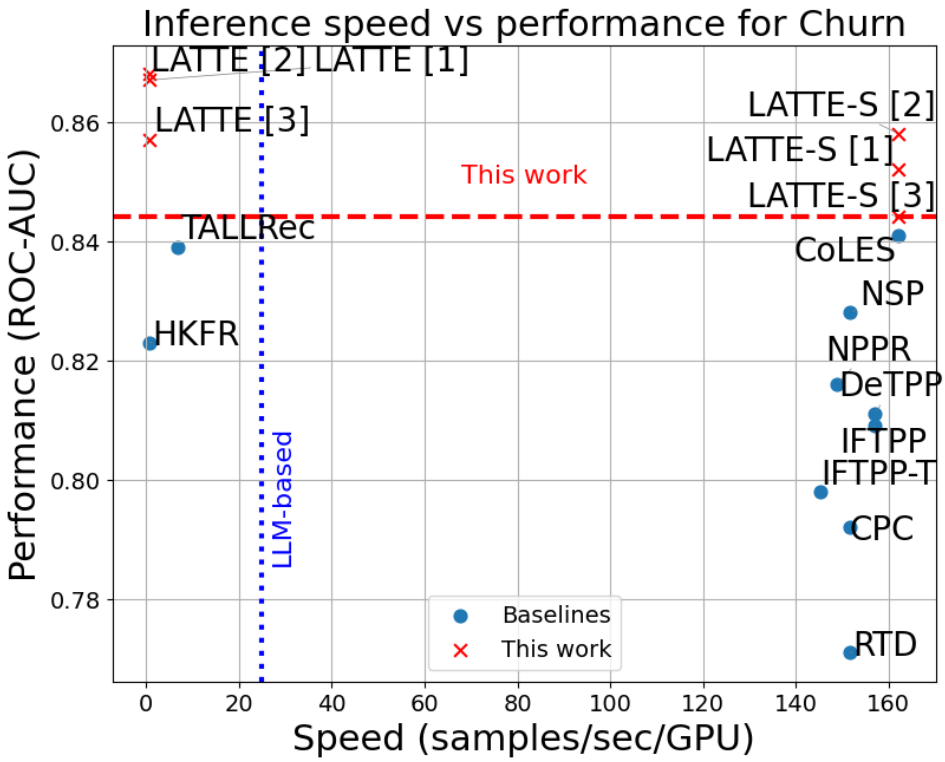

Figure 3: 表现优值图，比较了在客户流失数据集上模型在 ROC-AUC 和每 GPU 每秒样本推断速度方面的性能。

基于文本嵌入对齐的对比微调。 表 2 总结了一项关于对比微调和模态连接的消融研究。添加 LLM 生成的描述一致地提升了所有头部的性能——例如，**LATTE** [2] 在流失上的 AUC 从 0.858 提升到 0.868，在性别上的 AUC 从 0.894 提升到 0.901。对比对齐被证明是必不可少的，因为所有 **LATTE** 变体的性能都优于仅依赖特征连接的 CoLES + 描述。**LATTE** [3] 在年龄 (0.678) 和性别 (0.902) 上取得了最佳结果，突显了正则化的影响。这些结果表明，通过对比监督进行的语言定位优于简单的融合。

用于描述生成的语言模型选择。 表格 3 评估了为行为描述生成选择的大型语言模型（LLM）如何影响下游任务性能。在所有的对比头中，较大的指令调优模型始终提供更强的结果。特别是，Qwen 3 32B 在所有头上的用户流失预测中表现优于较小的替代品，在 **LATTE** [1] 下达到最佳的 0.862 AUC 得分。同时，Gemma 3 27B 在年龄和性别分类中提供了竞争力的表现，在 **LATTE** [2] (0.671 / 0.901) 和 **LATTE** [3] (0.678 / 0.902) 下均获得最高分。相比之下，最小的模型，Gemma 3 4B，在所有环境下均表现落后。这些结果强调了大型语言模型的生成能力对于高质量表示学习的重要性。

Contrastive Head	LLM Generator	Churn (AUC)	Age Group (Acc)	Gender (AUC)
LATTE [1]	Gemma 3 4B	0.849 ś 0.005	0.645 ś 0.006	0.888 ś 0.005
LATTE [1]	Gemma 3 27B	0.867 ś 0.004	0.666 ś 0.005	0.900 ś 0.005
LATTE [1]	Qwen 3 32B	0.859 ś 0.004	0.658 ś 0.005	0.896 ś 0.003
LATTE [2]	Gemma 3 4B	0.851 ś 0.005	0.652 ś 0.006	0.889 ś 0.002
LATTE [2]	Gemma 3 27B	0.868 ś 0.006	0.671 ś 0.005	0.901 ś 0.005
LATTE [2]	Qwen 3 32B	0.862 ś 0.001	0.667 ś 0.004	0.896 ś 0.000
LATTE [3]	Gemma 3 4B	0.849 ś 0.005	0.665 ś 0.006	0.890 ś 0.005
LATTE [3]	Gemma 3 27B	0.857 ś 0.003	0.678 ś 0.003	0.902 ś 0.006
LATTE [3]	Qwen 3 32B	0.862 ś 0.004	0.675 ś 0.002	0.895 ś 0.003

Table 3: 消融实验：选择语言模型生成行为描述的效果。

Contrastive Head	Text Encoder	Churn (AUC)	Age Group (Acc)	Gender (AUC)
LATTE [1]	mE5-large-instruct	0.860 ś 0.005	0.658 ś 0.008	0.895 ś 0.003
LATTE [1]	Qwen3-Embedding-0.6B	0.862 ś 0.004	0.660 ś 0.007	0.897 ś 0.003
LATTE [1]	Qwen3-Embedding-8B	0.867 ś 0.004	0.666 ś 0.005	0.900 ś 0.005
LATTE [2]	mE5-large-instruct	0.860 ś 0.007	0.664 ś 0.006	0.894 ś 0.004
LATTE [2]	Qwen3-Embedding-0.6B	0.863 ś 0.005	0.667 ś 0.005	0.896 ś 0.003
LATTE [2]	Qwen3-Embedding-8B	0.868 ś 0.006	0.671 ś 0.005	0.901 ś 0.005
LATTE [3]	mE5-large-instruct	0.850 ś 0.004	0.671 ś 0.003	0.897 ś 0.002
LATTE [3]	Qwen3-Embedding-0.6B	0.852 ś 0.006	0.673 ś 0.001	0.899 ś 0.002
LATTE [3]	Qwen3-Embedding-8B	0.857 ś 0.003	0.678 ś 0.003	0.902 ś 0.006

Table 4: 消融研究：文本嵌入模型对下游任务性能的影响。

文本嵌入模型选择。 表 4 探讨了用于获得 z_i^{text} 的冻结文本编码器对后续任务性能的影响，其中生成的嵌入既用于对齐也作为世界知识的独立来源。我们将 mE5-large-instruct [44] 与 Qwen3 [45] 的两个变体进行比较：一个轻量级的嵌入模型（0.6B）和一个大型的嵌入模型（8B）。在所有对比头中，Qwen3-Embedding-8B 一贯表现出最强的结果，年龄和性别任务中表现比 mE5 高 1.0-1.3 %，在用户流失任务中高 0.7 %。值得注意的是，在 **LATTE** [3] 下，它将年龄分类提升至 0.678，性别 AUC 提升至 0.902，二者均为最新的技术水平。附录 B 提供了更多关于不同描述嵌入策略比较的消融研究。

Contrastive Head	Prompt Format	Churn (AUC)	Age Group (Acc)	Gender (AUC)
LATTE [1]	Raw serialization	0.854 ś 0.003	0.653 ś 0.007	0.887 ś 0.004
LATTE [1]	Stats	0.867 ś 0.004	0.666 ś 0.005	0.900 ś 0.005
LATTE [2]	Raw serialization	0.855 ś 0.005	0.656 ś 0.007	0.886 ś 0.005
LATTE [2]	Stats	0.868 ś 0.006	0.671 ś 0.005	0.901 ś 0.005
LATTE [3]	Raw serialization	0.842 ś 0.007	0.662 ś 0.005	0.889 ś 0.003
LATTE [3]	Stats	0.857 ś 0.003	0.678 ś 0.003	0.902 ś 0.006

Table 5: 消融实验：输入格式为提示 LLM 原始交易序列化与汇总统计数据。

提示输入格式。 表 5 比较了提示 LLM 的两种格式：原始序列化交易序列和结构化行为摘要。跨所有对比头，使用统计摘要始终优于原始序列化。例如，在 **LATTE** [3] 设置中，统计提示将用户流失的 ROC-AUC 提高了 1.5 % 和性别 AUC 提高了 1.3 % 。附录 B 提供了文本描述效果相对于原始统计特征的额外消融分析。

我们提出了一种用于对比表示学习的创新方法，该方法利用合成生成的文本描述作为补充模态来从事件序列中学习。通过将结构化的交易数据与由冻结的经过指令调整的大型语言模型生成的自然语言总结对齐，该方法在嵌入空间中引入了文本先验，而无需标记数据或对大型语言模型进行微调。

LATTE 在三个关键的开源银行任务上达到了最新的成果，相比基线，性别预测提升了 6.1 %，年龄组分类提升了 3.0 %，流失预测提升了 2.0 %。所提出的 **LATTE**-S 在资源使用上非常高效（4.4M 参数，速度可达每秒 200 个样本），这对于行为日志丰富但监督有限的工业应用而言是必不可少的。

一个特别有前景的未来方向是探索更丰富的文本到序列对齐形式，其中自然语言摘要与潜在的事件动态结合得更加紧密。这种方法可以产生不仅对于预测任务有效而且本质上易于解释的嵌入，因为它们仍然与人类可读的用户行为描述紧密联系。

5

局限性

虽然我们的方法展示了强劲的经验表现，但它依赖于一个假设，即 LLM 生成的文本描述能够准确反映结构化序列中存在的行为细微差别。因此，冻结的 LLM 的提示设计和泛化能力限制了监督的质量。此外，文本编码器在训练中未更新，这可能会在分布偏移的领域中限制对齐的准确性。最后，尽管我们的方法不需要标签或微调，但与轻量级对比学习方法相比，由于依赖于大规模基于 LLM 的生成，该方法引入了适度的训练开销。

References

[1] Dmitrii Babaev, Nikita Ovsov, Ivan Kireev, Maria Ivanova, Gleb Gusev, Ivan Nazarov, and Alexander Tuzhilin. CoLES: Contrastive Learning for Event Sequences with Self-Supervision . In Proceedings of the 2022 International Conference on Management of Data , pages 1190–1199, 2022.

[2] Rosbank Churn Prediction Challenge . https://boosters.pro/championship/rosbank1/overview, 2021. Accessed: 2025-07-04.

[3] Python and Analyze Data: Final Project (Gender) . https://www.kaggle.com/competitions/python-and-analyze-data-final-project, 2021. Accessed: 2025-07-04.

[4] Sberbank Sirius Age Group Competition . https://ods.ai/competitions/sberbank-sirius-lesson, 2021. Accessed: 2025-07-04.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations . In Proceedings of the 37th International Conference on Machine Learning (ICML 2020) , Paper 149, 11 pages.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark et al. Learning Transferable Visual Models from Natural Language Supervision . In International Conference on Machine Learning , pages 8748–8763, 2021.

[7] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language–Image Pre-training . In Proceedings of the IEEE/CVF International Conference on Computer Vision , pages 11975–11986, 2023.

[8] Nikita Severin, Aleksei Ziablitsev, Yulia Savelyeva, Valeriy Tashchilin, Ivan Bulychev, Mikhail Yushkov, Artem Kushneruk, Amaliya Zaryvnykh, Dmitrii Kiselev, Andrey Savchenko, and Ilya Makarov. LLM-KT: A Versatile Framework for Knowledge Transfer from Large Language Models to Collaborative Filtering . arXiv preprint arXiv:2411.00556 , 2024.

[9] Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. TIMECMA: Towards LLM-Empowered Multivariate Time-Series Forecasting via Cross-Modality Alignment . In Proceedings of the AAAI Conference on Artificial Intelligence , 39(18):18780–18788, 2025.

[10] Shitong Dai, Jiongnan Liu, Zhicheng Dou, Haonan Wang, Lin Liu, Bo Long, and Ji-Rong Wen. Contrastive Learning for User Sequence Representation in Personalized Product Search . In KDD ' 23: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining , pages 380–389, 2023.

[11] Trisha Das, Zifeng Wang, and Jimeng Sun. TWIN: Personalized Clinical Trial Digital Twin Generation . In KDD ' 23: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining , pages 402–413, 2023.

[12] Zitao Liu, Qiongqiong Liu, Teng Guo, Jiahao Chen, Shuyan Huang, Xiangyu Zhao, Jiliang Tang, Weiqi Luo, and Jian Weng. XES3G5M: A Knowledge Tracing Benchmark Dataset with Auxiliary Information . In Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track , 2023.

[13] Dzhambulat Mollaev, Alexander Kostin, Maria Postnova, Ivan Karpukhin, Ivan Kireev, Gleb Gusev, and Andrey Savchenko. Multimodal Banking Dataset: Understanding Client Needs through Event Sequences . arXiv preprint arXiv:2409.17587 , 2025.

[14] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A Survey on Self-Supervised Learning: Algorithms, Applications, and Future Trends . IEEE Transactions on Pattern Analysis and Machine Intelligence , 2024.

[15] Irina Abdullaeva, Andrei Filatov, Mikhail Orlov, Ivan Karpukhin, Viacheslav Vasilev, Denis Dimitrov, Andrey Kuznetsov, Ivan Kireev, and Andrey Savchenko. ESQA: Event Sequences Question Answering . arXiv preprint arXiv:2407.12833 , 2024.

[16] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan et al. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models . In Proceedings of the 12th International Conference on Learning Representations (ICLR) , 2024.

[17] Chenxi Sun, Hongyan Li, Yaliang Li, and Shenda Hong. TEST: Text Prototype-Aligned Embedding to Activate LLM ' s Ability for Time Series . In Proceedings of the 12th International Conference on Learning Representations (ICLR) , 2024.

[18] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. TabLLM: Few-Shot Classification of Tabular Data with Large Language Models . In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS) , pages 5549–5581, 2023.

[19] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. TallRec: An Effective and Efficient Tuning Framework to Align Large Language Models with Recommendation . In Proceedings of the 17th ACM Conference on Recommender Systems , pages 1007–1014, 2023.

[20] Zhi Zheng, Wen-Shuo Chao, Zhaopeng Qiu, Hengshu Zhu, and Hui Xiong. Harnessing Large Language Models for Text-Rich Sequential Recommendation . In Proceedings of The Web Conference 2024 , 2024.

[21] Bin Yin, Junjie Xie, Yu Qin, Zixiang Ding, Zhichao Feng, Xiang Li, and Wei Lin. Heterogeneous Knowledge Fusion: A Novel Approach for Personalized Recommendation via LLM . In Proceedings of the 17th ACM Conference on Recommender Systems , pages 599–601, 2023.

[22] Piotr Skalski, David Sutton, Stuart Burrell, Iker Perez, and Jason Wong. Towards a Foundation Purchasing Model: Pretrained Generative Autoregression on Transaction Sequences . In Proceedings of the 4th ACM International Conference on AI in Finance , pages 141–149, 2023.

[23] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding . arXiv preprint arXiv:1807.03748 , 2018.

[24] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-Training Text Encoders as Discriminators Rather Than Generators . In International Conference on Learning Representations , 2019.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding . In Proceedings of NAACL-HLT 2019 , pages 4171–4186.

[26] Igor Udovichenko, Egor Shvetsov, Denis Divitsky, Dmitry Osin, Ilya Trofimov, Ivan Sukharev, Anatoliy Glushenko, Dmitry Berestnev, and Evgeny Burnaev. SeqNAS: Neural Architecture Search for Event Sequence Classification . IEEE Access , 12:3898–3909, 2024.

[27] Anton Yeshchenko and Jan Mendling. A Survey of Approaches for Event Sequence Analysis and Visualization using the ESeVis Framework . arXiv preprint arXiv:2202.07941 , 2022.

[28] Bojan Kolosnjaji, Apostolis Zarras, George Webster, and Claudia Eckert. Deep Learning for Classification of Malware System Call Sequences . In AI 2016: Advances in Artificial Intelligence, 29th Australasian Joint Conference , pages 137–149.

[29] Gary M. Weiss and Haym Hirsh. Learning to Predict Rare Events in Event Sequences . In Proceedings of KDD 1998 , pages 359–363.

[30] Yi Guo, Shunan Guo, Zhuochen Jin, Smiti Kaul, David Gotz, and Nan Cao. Survey on Visual Analysis of Event Sequence Data . arXiv preprint arXiv:2006.14291 , 2020.

[31] Dmitrii Babaev, Maxim Savchenko, Alexander Tuzhilin, and Dmitrii Umerenkov. ET-RNN: Applying Deep Learning to Credit Loan Applications . In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , pages 2183–2190, 2019.

[32] Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and Constructing Latent Modality Structures in Multi-Modal Representation Learning . In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , pages 7661–7671, 2023.

[33] Ivan Karpukhin and Andrey Savchenko. DeTPP: Leveraging Object Detection for Robust Long-Horizon Event Prediction . arXiv preprint arXiv:2408.13131 , 2024.

[34] Oleksandr Shchur, Marin Bilo, and Stephan Günnemann. Intensity-Free Learning of Temporal Point Processes . In International Conference on Learning Representations , 2020.

[35] Simone Luetto, Fabrizio Garuti, Enver Sangineto, Lorenzo Forni, and Rita Cucchiara. One Transformer for All Time Series: Representing and Training with Time-Dependent Heterogeneous Tabular Data . Machine Learning , 114(6):1–27, 2025.

[36] Yue Wang, Tianfan Fu, Yinlong Xu, Zihan Ma, Hongxia Xu, Bang Du, Yingzhou Lu, Honghao Gao, Jian Wu, and Jintai Chen. TWIN-GPT: Digital Twins for Clinical Trials via Large Language Model . ACM Transactions on Multimedia Computing, Communications and Applications , 2024.

[37] Jiongnan Liu, Zhicheng Dou, Jian-Yun Nie, Zhenlin Chen, Guoyu Tang, Sulong Xu, and Ji-Rong Wen. Enhancing Sequential Personalized Product Search with External Out-of-Sequence Knowledge . ACM Transactions on Information Systems , 43(4):1–25, 2025.

[38] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree . Advances in Neural Information Processing Systems 30 , 2017.

[39] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard et al. Gemma 3 Technical Report . arXiv preprint arXiv:2503.19786 , 2025.

[40] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang et al. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models . arXiv preprint arXiv:2506.05176 , 2025.

[41] Dmitry Osin, Igor Udovichenko, Viktor Moskvoretskii, Egor Shvetsov, and Evgeny Burnaev. EBES: Easy Benchmarking for Event Sequences . arXiv preprint arXiv:2410.03399 , 2025.

[42] Niklas Muennighoff, Su Hongjin, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative Representational Instruction Tuning . In ICLR 2024 Workshop: How Far Are We From AGI , 2024.

[43] Yucheng Ruan, Xiang Lan, Jingying Ma, Yizhi Dong, Kai He, and Mengling Feng. Language Modeling on Tabular Data: A Survey of Foundations, Techniques and Evolution . arXiv preprint arXiv:2408.10548 , 2024.

[44] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual E5 Text Embeddings: A Technical Report . arXiv preprint arXiv:2402.05672 , 2024.

[45] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng et al. Qwen3 Technical Report . arXiv preprint arXiv:2505.09388 , 2025.

[46] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders . In First Conference on Language Modeling , 2024.

[47] Hugo Touvron et al. The LLaMA 3 Herd of Models . arXiv preprint arXiv:2407.21783 , 2024.

[48] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. GLM: General Language Model Pretraining with Autoregressive Blank Infilling . In Proceedings of ACL 2022 (long papers) , pages 320–335.

[49] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux et al. LLaMA: Open and Efficient Foundation Language Models . arXiv preprint arXiv:2302.13971 , 2023.

A 实验细节

对于所有实验，使用经指令调整的 `Gemma-3-27B-Instruct` [39] 模型生成自然语言描述。为了将生成的行为描述编码为固定长度的向量，我们采用了 `Qwen3-Embedding-8B` [40] 模型，这是一个针对语义检索优化的多语言编码器。事务编码器被实例化为在 CoLES 自监督框架下训练的基于 GRU 的模型，作为所有对比对齐头的基础序列编码器。整个管道的训练和推理，包括 LLM 提示、嵌入计算和对比对齐，均在单个 NVIDIA Tesla A100 80GB GPU 上执行。

B 附加实验

Contrastive Head	Auxiliary Modality	Churn (AUC)	Age Group (Acc)	Gender (AUC)
LATTE [1]	Stats (MLP-1L)	0.846 ś 0.006	0.647 ś 0.006	0.886 ś 0.004
LATTE [1]	Stats (MLP-2L)	0.847 ś 0.005	0.645 ś 0.007	0.887 ś 0.003
LATTE [1]	LLM text	0.867 ś 0.004	0.666 ś 0.005	0.900 ś 0.005
LATTE [2]	Stats (MLP-1L)	0.848 ś 0.005	0.651 ś 0.006	0.883 ś 0.005
LATTE [2]	Stats (MLP-2L)	0.851 ś 0.003	0.653 ś 0.005	0.884 ś 0.004
LATTE [2]	LLM text	0.868 ś 0.006	0.671 ś 0.005	0.901 ś 0.005
LATTE [3]	Stats (MLP-1L)	0.842 ś 0.007	0.657 ś 0.004	0.886 ś 0.003
LATTE [3]	Stats (MLP-2L)	0.843 ś 0.005	0.656 ś 0.003	0.887 ś 0.003
LATTE [3]	LLM text	0.857 ś 0.003	0.678 ś 0.003	0.902 ś 0.006

Table 6: 消融研究：用于对齐的辅助模态的效果——手工制作的统计数据与 LLM 生成的文本。粗体表示最佳，每组中 underline 代表次佳。

文本嵌入与统计嵌入 表格 6 比较了两种辅助监督目标：通过 MLP 投影的原始统计特征向量与 LLM 生成的自然语言描述。基于文本的监督在每个对比头的所有任务中都表现出更强的下游性能。这种优势在 **LATTE** [3] 下尤其明显，其中基于 LLM 的监督将年龄分类准确性提高到 0.678，性别 AUC 提高到 0.902（整体最高）。即使是表现最佳的 2 层 MLP 投影在大多数情况下也逊色 1–2 %。这些结果证实，自然语言描述比手工制作的统计数据为序列表示学习提供了更具信息量的目标。

Contrastive Head	Method	Churn (AUC)	Age Group (Acc)	Gender (AUC)
LATTE [1]	LLM encoder (ours)	0.867 ś 0.004	0.666 ś 0.005	0.900 ś 0.005
LATTE [1]	MeanPool (generator)	0.859 ś 0.006	0.659 ś 0.003	0.893 ś 0.001
LATTE [2]	LLM encoder (ours)	0.868 ś 0.006	0.671 ś 0.005	0.901 ś 0.005
LATTE [2]	MeanPool (generator)	0.862 ś 0.005	0.666 ś 0.002	0.894 ś 0.003
LATTE [3]	LLM encoder (ours)	0.857 ś 0.003	0.678 ś 0.003	0.902 ś 0.006
LATTE [3]	MeanPool (generator)	0.851 ś 0.007	0.671 ś 0.004	0.894 ś 0.002

Table 7: 消融：对比从专用句子编码器（我们的）获得的描述嵌入与直接从生成器 LLM 获得的 的区别。

文本嵌入提取策略的影响。 我们比较了从行为描述中获取文本嵌入的两种策略。在我们的默认设置中，我们使用一个冻结的句子编码器（Qwen3-Embedding-8B）来计算嵌入，将生成和编码过程分开。作为替代方法，我们

通过对隐藏状态进行均值池化，直接从生成器 LLM（Gemma-3-27B）中提取嵌入。根据最近研究表明，对所有标记嵌入进行均值池化优于使用 EOS 标记 [46, 42]，我们在最后 $k = 8$ 个变换器层上对隐藏状态取平均。举例来说，使用专用的 LLM 编码器比直接从生成器进行池化在性别和流失率 AUC 方面分别提高了最多 0.8 %。

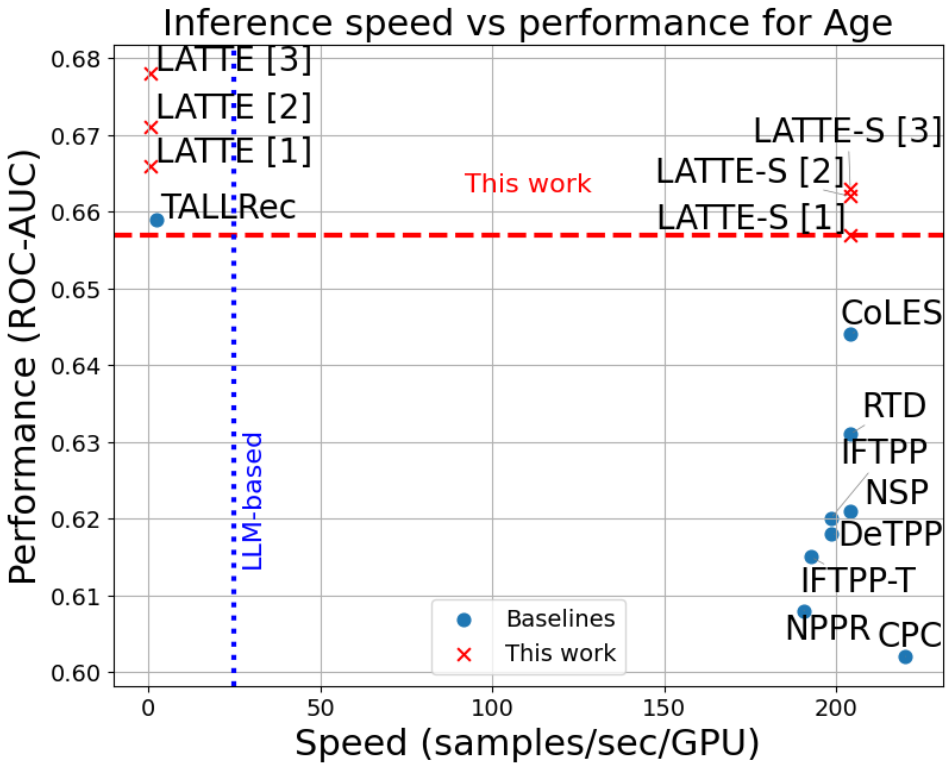

Figure 4: 模型在 ROC-AUC 上的表现和在 Age 数据集上每 GPU 每秒样本数的推理速度的评价指标比较。

运行时间分析 图 4 显示了年龄预测任务的结果。我们观察到与流失数据集（图 3）一致的趋势：生成模型达到最高的准确性（最高达 0.678），但其推理吞吐量有限。相反，轻量级的 LATTE-S 变体达到了良好的权衡，维持了具竞争力的性能（约 0.66 ROC-AUC），同时显著提高了推理速度，超过每秒 200 个样本。

C 文本生成协议

我们通过提示一个经过指令微调的大型语言模型，从统计用户摘要生成自然语言描述。表格 8 展示了用于行为描述生成的完整提示，而表格 9 提供了模型生成的输出示例。

System: You are an expert in financial transaction analysis. Your task is to generate clear, structured, and concise descriptions of user financial behavior based on given transaction data. Use data-driven insights and avoid speculation.

Guidelines:

- 不要包含诸如“这是用户 X 的财务行为描述”这样的短语
- 直接从行为洞察开始
- 将数字解释为模式（例如，爆发性支出，常规支付）
- 突出消费习惯、风险因素、财务一致性
- 避免不支持的假设；保持清晰和简洁

User: Below is a summary of a user’s transaction history:

```
- Number of transactions: 226
- Active transaction period: 444 days
- Avg transactions per day: 0.51
- Avg transaction interval: 1.97 days
- Top MCCs: Supermarkets, cash withdrawals, telecommunications
- Share of transactions in Top 1 MCC: 0.04
- Avg amount for Top 1 MCC: 2770.47 RUB
- Avg amount for Top 2 MCC: 45969.41 RUB
- Avg amount for Top 3 MCC: 8139.52 RUB
- Share of days with transactions: 0.36
- Total income: 2,706,553.10 RUB
- Total expenses: 3,956,007.97 RUB
- Avg outgoing amount: 17,981.85 RUB
- Avg incoming amount: 451,092.18 RUB
```

Instructions:

- 分析行为模式
- 识别交易规律性和类别依赖性
- 评估潜在风险因素和财务规划特征
- 在保持事实准确的同时，以结构化和引人入胜的方式撰写

Table 8: 行为描述生成的提示模板

Output:
This user demonstrates a low-frequency, high-variance spending pattern over a substantial period (444 days). Transactions occur on approximately 36 % of days, averaging one every two days, indicating infrequent but potentially substantial financial activity.
Spending is heavily weighted towards cash withdrawals from financial institutions (4 % of transactions, average 45969.41 RUB), suggesting a preference for cash-based expenses or large withdrawals for specific purposes. Significant spending also occurs at grocery stores/supermarkets (average 8139.52 RUB), representing a consistent, though smaller, portion of overall expenditure. A smaller number of transactions are related to phone services (average 2770.47 RUB).
Expenses significantly exceed income (3,956,007.97 RUB vs. 2,706,553.10 RUB), implying reliance on savings, credit, or other income sources not captured in this data. The high average outgoing transaction amount (17,981.85 RUB) coupled with a large average incoming amount (451,092.18 RUB) suggests irregular but substantial income events are occurring, potentially offsetting consistent overspending.
The combination of infrequent transactions, high cash withdrawals, and a substantial expense-to-income imbalance points to potentially inconsistent financial management . While not indicative of reckless spending, the reliance on infrequent income and high cash usage warrants further investigation to determine financial stability and potential risk factors. There ' s a lack of diversification in spending, with the top three MCCs dominating activity.

Table 9: 从 LLM 生成的示例输出