

大语言模型增强的时间思维链图书推荐模型^①



胡志臣¹, 王 明¹, 酆子莹², 赵永杰¹

¹(镇江市高等专科学校 图文信息中心, 镇江 212028)

²(镇江市高等专科学校 学生处, 镇江 212028)

通信作者: 酆子莹, E-mail: ziyingli@zjc.edu.cn

摘 要: 在信息过载的时代, 精准的图书推荐对于提高用户的阅读体验, 发掘优质内容有重要的意义. 但是现有的推荐系统都存在着动态兴趣捕捉能力弱、图书深层语义理解不足和推荐可解释性差的问题. 因此本文提出一种基于大语言模型增强的时间思维链图书推荐模型 (Time-CoT-LLM). 该模型用时间思维链模块结合用户近期行为动态来形成语义推理链, 加强推荐的可解释性. 使用融合时间稀疏注意力 (time-invasive sparse attention, TISA) 与混合专家 (mixture-of-experts, MoE) 的用户编码器来精准捕捉用户兴趣随时间的变化及多样性, 借助 LLM 增强的图书编码器挖掘图书深层语义特征. 在 BookCrossing 和 Goodreads 两个数据集上的实验结果表明 Time-CoT-LLM 在 *AUC* (0.772)、*NDCG@5* (0.396) 等指标上比传统的协同过滤、BERT-based 序列推荐、纯 LLM 模型提高了 3.2%–4.1%, 85% 的用户认为推荐理由更清晰. 本研究为图书推荐中的动态兴趣、语义理解、可解释性问题提供了有效的解决途径, 对智能推荐技术在文化领域的落地有重要的推动作用.

关键词: 大语言模型; 思维链; 图书推荐; 时间稀疏注意力; 语义推理

引用格式: 胡志臣, 王明, 酆子莹, 赵永杰. 大语言模型增强的时间思维链图书推荐模型. 计算机系统应用. <http://www.c-s-a.org.cn/1003-3254/10226.html>

Time Chain-of-thought Book Recommendation Model Enhanced by Large Language Model

HU Zhi-Chen¹, WANG Ming¹, LI Zi-Ying², ZHAO Yong-Jie¹

¹(Library and Information Center, Zhenjiang College, Zhenjiang 212028, China)

²(Student Affairs Office, Zhenjiang College, Zhenjiang 212028, China)

Abstract: In an era of information overload, precise book recommendations are of significance for improving users' reading experiences and discovering high-quality content. However, existing recommendation systems suffer from problems including the weak ability to capture dynamic interests, insufficient deep semantic understanding of books, and poor interpretability of recommendations. To this end, this study proposes a time chain-of-thought book recommendation model enhanced by large language models (Time-CoT-LLM). This model employs a time chain-of-thought module to integrate users' recent behavioral dynamics, forming semantic inference chains and enhancing recommendation interpretability. A user encoder combining time-invasive sparse attention (TISA) with mixture-of-experts (MoE) is adopted to precisely capture the temporal variation and diversity of user interests. An LLM-enhanced book encoder is utilized to extract deep semantic features from books. Experiments conducted on the BookCrossing and Goodreads datasets demonstrate that Time-CoT-LLM outperforms traditional collaborative filtering, BERT-based sequence recommendation, and pure LLMs by 3.2% to 4.1% in metrics such as *AUC* (0.772) and *NDCG@5* (0.396). Furthermore, 85% of the users perceive the recommendation reasons as clearer. This study provides an effective solution for dealing

① 基金项目: 国家自然科学基金 (62202210); 安徽省教育厅优秀青年基金 (YQYB202316); 镇江市高等专科学校校级科研课题基金 (GZYB202508)

收稿时间: 2025-12-30; 修改时间: 2026-01-19, 2026-01-30, 2026-02-09; 采用时间: 2026-02-13; csa 在线出版时间: 2026-06-18

with challenges in dynamic interest tracking, semantic understanding, and interpretability in book recommendations, and plays a significant role in advancing the practical application of intelligent recommendation technologies in cultural domains.

Key words: large language model (LLM); chain-of-thought (CoT); book recommendation; time-invasive sparse attention (TISA); semantic reasoning

1 引言

数字阅读时代到来以后,图书资源数量急剧增加与用户选择能力有限之间的矛盾愈加突出^[1].精准的图书推荐系统可以快速帮助用户找到自己感兴趣的好书,提高用户的阅读体验,同时也有利于优质图书的传播,发掘内容的价值^[2].但是传统的推荐方法仍然存在很多不足,协同过滤依靠用户和图书的交互矩阵,不能很好地解决冷启动和数据稀疏的问题.基于内容的推荐依靠手工特征,不能够体现图书的深层语义.序列推荐用 LSTM、Transformer 编码行为序列,但是对时间权重的处理比较简单,不能很好地捕捉用户兴趣的动态变化^[3].

近年来,以 BERT、GPT-3.5、LLaMA 为代表的大语言模型 (large language model, LLM) 的出现给推荐系统带来了新的机会^[4-6]. LLM 具有较强的语义理解能力,可以从图书标题、摘要、评论等文本中挖掘出深层次的语义特征^[7].思维链技术的使用令 LLM 可以产生逐步推理的过程.由此可以看出,LLM 可以解决传统推荐方法的不足,给图书推荐系统的升级提供了新的思路^[8].

然而,现有的推荐系统仍然存在 3 大主要问题:(1) 动态兴趣捕捉能力不足,用户阅读兴趣会随着时间的推移而发生变化,但是现有的模型大多使用固定的权重来处理历史行为,不能很好地捕捉到这种变化^[9];(2) 图书深层语义理解不足,图书的主题、风格等细粒度特征需要结合类别、用户评论等多源信息,现有的 LLM 编码只依赖文本内容,无法区分硬科幻和软科幻、历史小说和传记等类别;(3) 推荐可解释性差,纯 LLM 模型的推荐理由都是生成的自然语言,并没有结合用户的实际行为,用户无法信任推荐结果^[10].

为了解决上述问题,本文提出一种基于大语言模型增强的时间思维链图书推荐模型 Time-CoT-LLM,将 LLM 的语义推理能力与时间动态建模结合起来,实

现准确、自动、可解释的图书推荐.模型主要由 4 个模块组成,分别是图书编码器模块、时间思维链模块、用户编码器模块、融合推荐模块.使用 DistilBERT 对图书标题、摘要、目录等特征进行编码,同时将时间思维链编码为语义向量,将两者融合得到图书的深层语义表示.用户编码器用时间稀疏注意力 (time-invasive sparse attention, TISA) 计算历史行为的时间权重,用混合专家 (mixture-of-experts, MoE) 模块处理兴趣多样性,输出用户的动态兴趣向量.融合推荐模块,计算用户兴趣向量与图书语义向量的余弦相似度,排序后输出推荐列表,并附上时间思维链作为推荐理由.本文的主要贡献如下.

(1) 提出时间思维链生成机制,结合用户近期行为动态调整推理链的语义重点,解决了现有思维链静态化的问题,显著提升了推荐的可解释性.

(2) 设计融合时间稀疏注意力与混合专家的用户编码器,精准捕捉用户兴趣的时间演化与多样性,解决了现有模型动态兴趣捕捉能力弱的问题.

(3) 构建 LLM 增强的图书编码器,融合图书内容与时间思维链的语义特征,提升了图书深层语义理解能力.

本文第 2 节对传统图书推荐方法以及 LLM 在推荐中应用建模的相关工作进行综述.第 3 节给出 Time-CoT-LLM 的整体架构以及各模块的设计.第 4 节通过实验验证模型的性能,并和基线模型进行对比分析.第 5 节总结研究结果,提出不足与今后的发展方向.

2 相关工作

2.1 传统图书推荐方法

个性化推荐是高校智慧图书馆建设的重要一环,但是传统的方法普遍存在着用户画像不够精确、冷启动、兴趣融合不足、结构信息利用不充分等问题,导致推荐不能满足用户真正的需求.从不同角度提出改

进方案,促进传统图书推荐走向智能化。

对于用户画像和推荐算法融合不够的问题,王大阜等人^[11]提出了用户画像结合混合推荐的系统架构,从读者属性、行为、兴趣等方面建立用户画像,划分身份类型;然后结合协同过滤、内容推荐、属性相似度的混合算法,基于 Hadoop 平台推荐 Top- N 图书,实现了用户特征与推荐策略的有效匹配。冷启动一直是推荐的经典难题,Ahmed 等人^[12]提出了一种矩阵分解协同滤波的方法,将用户和物品的交互矩阵分解成两个低秩矩阵,挖掘用户的潜在兴趣以及图书的特征,从而解决新用户和新图书的推荐问题,并且比较了基于内存的邻域法和基于模型的矩阵分解法,在冷启动场景下验证了基于模型的方法效果更好。

为解决用户长期与短期兴趣融合不充分的问题,易晓宇等人^[13]提出聚类结合兴趣融合方法,用 K 均值聚类优化用户分组,计算兴趣度确定候选资源;再通过 LSKGCN 模型融合长期与短期兴趣,通过评分机制推荐高相关性图书,实现了兴趣的动态融合。针对图书内容结构信息未充分利用的问题,王莉军等人^[14]构建“图结构+相似计算”方法,对图书文本进行短语抽取,用 TextRank 提取关键词,建立图书向量;结合 HNSW 算法计算相似度,有效利用了图书内容的结构特征,提升了相似推荐的准确性。

为增强情感分析在推荐中的作用,Devika 等人^[15]引入情感分析结合集成学习模型,通过 TextBlob 评估亚马逊评论极性,用 BERT 生成语义嵌入;通过集成学习融合多模型输出,增强了推荐系统的性能与精度,实现了情感信息与推荐的有效结合。

2.2 LLM 推荐应用方法

基于文本的推荐由于可以推荐所有物品而被广泛使用,但是原始的商品描述中缺少与用户偏好的全面相关的信息,导致推荐不够准确。LLM 的语义理解与常识推理能力为解决这个问题提供了机会,但是应用过程中还存在着任务差异、效率瓶颈、公平性等问题。

针对原始文本缺乏用户偏好信息的问题,Lyu 等人^[16]提出 LLM-Rec 方法,融合 4 种文本丰富提示策略,增强文本的个性化信息,有效提升了文本推荐的精准度与个性化水平。由于 LLM 预训练任务与推荐任务差异显著,且预训练数据中缺乏推荐场景,其在推荐中的适应性有限。Lin 等人^[17]提出 TALLRec 调优框架,用推

荐数据对 LLM 进行针对性调校,构建大型推荐语言模型,将 LLM 的语义能力与推荐任务对齐,显著弥合了预训练与推荐的差距。

判别式推荐需逐条计算物品排名分数,效率低且不能处理复杂的上下文。Ji 等人^[18]提出了 GenRec 生成式推荐方法,利用 LLM 强大的表达能力直接生成推荐目标条目,而不是计算排名。通过解析用户行为上下文来学习用户的偏好,提高了推荐效率,并且具有较强的适应复杂场景的能力。大语言模型在推荐中的公平性问题越来越严重,已有的方法大多依靠显式的敏感属性来约束,但由于隐私保护,当属性缺失时就无法使用。龚镇辉等人^[19]提出了 FAIRSAA 框架,使用动态提示优化和对抗性重加权,无需敏感属性就可以提升公平性,从而解决缺少敏感属性而带来的公平性问题。

为融合 LLM 的语义理解与 GNN 的结构建模能力,解决推荐中语义与结构信息利用不足的问题,于恩海等人^[20]提出 LGSBR 会话推荐框架,用 LLM 生成商品补充文本嵌入与用户跨会话兴趣嵌入,融合文本与 ID 嵌入,结合对齐损失实现个性化推荐,提升了会话推荐的准确性。

结合传统图书推荐和 LLM 推荐的研究现状如表 1 所示,虽然各种方法在协同过滤、语义编码等各方面都有了很大的进步,但是还存在一些共性问题。许多方法在动态兴趣捕捉上表现不佳,传统序列推荐多采用固定权重处理历史行为,不能很好地捕捉到兴趣随时间的变化;即便加入时间信息,也没有充分考虑兴趣的多样性,造成推荐滞后。另外,目前的方法对图书的深层语义理解不够,LLM 增强的推荐方法虽然可以编码文本内容,但是没有融合用户行为相关的语义,不能区分硬科幻和软科幻这样的细粒度类别,造成语义相关性不够。还有推荐的可解释性差,纯 LLM 模型的推荐理由没有与用户的具体行为关联起来,推理链也没有根据用户的行为进行调整,所以用户无法信任。

针对目前方法中动态兴趣捕捉弱、图书深层语义理解不足、推荐可解释性差的问题,本文提出一种基于大语言模型增强的时间思维链图书推荐模型 Time-CoT-LLM,结合了大语言模型的语义推理能力和时间动态建模。该框架完成动态兴趣捕捉、深入语义理解、可解释推荐,给用户带来更好的服务。

表1 研究现状分析

文献	方法	目的	性能
文献[11]	用户画像+混合推荐系统架构(融合协同过滤、内容及属性相似度)	解决用户画像与推荐算法融合不足的问题	实现用户特征与推荐策略精准匹配,向目标读者推荐Top-N图书
文献[12]	矩阵分解协同滤波	克服冷启动问题	性能较传统方法有所提升
文献[13]	基于兴趣信息深度融合的网络图书资源推荐方法(K均值聚类+LSKGCN模型)	解决用户长期与短期兴趣融合不充分的问题	提升图书推荐的相关性与精准度
文献[14]	基于图结构的相似图书内容推荐方法(TextRank+HNSW)	解决图书内容结构信息未充分利用的问题	提升相似图书推荐的准确性
文献[15]	情感分析+集成学习模型(TextBlob+BERT+集成学习)	增强情感分析在推荐中的作用	增强推荐系统的性能与精度
文献[16]	LLM-Rec(4种文本丰富提示策略)	解决原始文本缺乏用户偏好信息的问题	提升文本推荐的精准度与个性化水平
文献[17]	TALLRec调优框架(用推荐数据调校LLM)	弥合LLM预训练与推荐任务的差距	显著提升LLM的推荐性能
文献[18]	GenRec生成式推荐方法	解决判别式推荐效率低、难以处理复杂上下文的问题	提升推荐效率与复杂场景适应能力
文献[19]	FAIRSAA框架(动态提示优化+对抗重加权)	解决敏感属性缺失场景下的推荐公平性问题	实现推荐公平性提升
文献[20]	LGSBR会话推荐框架(LLM生成嵌入+GNN融合)	融合LLM语义理解与GNN结构建模能力	提升会话推荐的准确性

3 方法

3.1 整体构建

本文提出 Time-CoT-LLM, 针对图书阅读场景的智能推荐模型, 在解决传统推荐系统“语义挖掘浅、时间动态缺失、可解释性不足”的痛点. 模型以图书语义增强、用户兴趣动态建模、推荐决策为核心逻辑, 整合4大模块: (1) 图书编码器模块, 融合内容与时间思维链语义; (2) 时间思维链模块, 动态生成用户个性化思维链; (3) 用户编码器模块, 捕捉兴趣时间演化与多样性; (4) 融合推荐模块, 生成可解释推荐.

如图1所示, 输入图书内容与用户历史行为, 通过图书编码器生成含时间语义的图书向量, 通过用户编码器生成含时间动态的用户向量, 最终计算相似度得到推荐列表, 并以时间思维链作为理由输出, 实现精准推荐和可理解性的双重优化.

3.2 图书编码器模块

图书编码器的主要目的是把图书客观内容特征和用户兴趣关联语义结合起来, 得到一个既有内容客观性又有用户个性化的图书语义向量. 模块包含内容特征提取、思维链生成、特征融合这3个主要步骤.

3.2.1 整体构建

针对图书标题、摘要、目录等结构化内容, 采用轻量化预训练语言模型 DistilBERT 进行语义编码, 捕捉图书主题、风格及核心元素等客观特征. 设图书集

合为 B , 对任意图书 $b \in B$, 其内容特征向量定义为, 如式(1)所示:

$$e_b^{\text{content}} = \text{DistilBERT}(\text{title}_b \oplus \text{abstract}_b \oplus \text{table}_b) \quad (1)$$

其中, title_b 、 abstract_b 、 table_b 分别为图书 b 的标题、摘要、目录文本, $\oplus$ 表示文本拼接; $e_b^{\text{content}} \in R^{dc}$, 为图书客观内容的语义表示.

3.2.2 时间思维链生成与编码

为增强图书语义的用户个性化, 引入时间思维链(Time-CoT), 基于用户近期阅读历史生成图书与用户兴趣匹配的推理链, 如“用户近期阅读《三体》, 目标图书属于硬科幻类别, 包含宇宙文明元素, 符合用户对宏大宇宙叙事的兴趣”. 设用户 u 的近期阅读历史子集为 H_u^{recent} , 取最近 K 本图书, 图书 b 的类别为 c_b , 则 Time-CoT 生成提示构造, 如图2所示.

用户 u 近期阅读图书的核心元素交集为 F_u^{inter} , 通过计算 H_u^{recent} 中所有图书核心元素的交集得到, 如式(2)所示:

$$F_u^{\text{inter}} = \bigcap_{h \in H_u^{\text{recent}}} (F_h) \quad (2)$$

其中, F_h 为图书 h 的核心元素集合, 由 DistilBERT 编码图书文本后, 通过关键词提取算法 (TextRank) 得到 Top-5 核心元素.

为确保推理链的时间敏感性, 引入时间衰减系数

α 对用户近期阅读行为的权重进行调整, 近期阅读图书在推理中的权重更高, 如式 (3) 所示:

$$w_h = \frac{e^{-\alpha \cdot \Delta t_h}}{\sum_{h' \in H_u^{\text{recent}}} e^{-\alpha \cdot \Delta t_{h'}}} \quad (3)$$

其中, w_h 为用户 u 阅读图书 h 的时间权重, Δt_h 为用户阅读图书 h 至当前的时间间隔 (单位: 天), α 取值为 0.1, 可使 30 天前的阅读行为权重衰减至近期 (1 天前) 的 4.5%, 90 天前权重衰减至 0.5%, 既突出近期行为的主导作用, 又保留远期行为的微弱影响, 符合用户兴趣演化规律. 将

时间权重融入提示词, 优化后的提示词如式 (4) 所示:

$$\text{Prompt}'(u, b) = \text{Prompt}(u, b) + \sum_{h \in H_u^{\text{recent}}} (w_h \times 100) + \text{title}_h \quad (4)$$

通过 ChatGLM2-6B 生成 Time-CoT 的推理文本 $\text{CoT}(u, b)$, 如式 (5) 所示:

$$\text{CoT}(u, b) = \text{ChatGLM2-6B}(\text{Prompt}'(u, b)) \quad (5)$$

其中, $\text{ChatGLM2-6B}(\cdot)$ 表示 ChatGLM2-6B 的文本生成函数, 输出为自然语言格式的推理链文本.

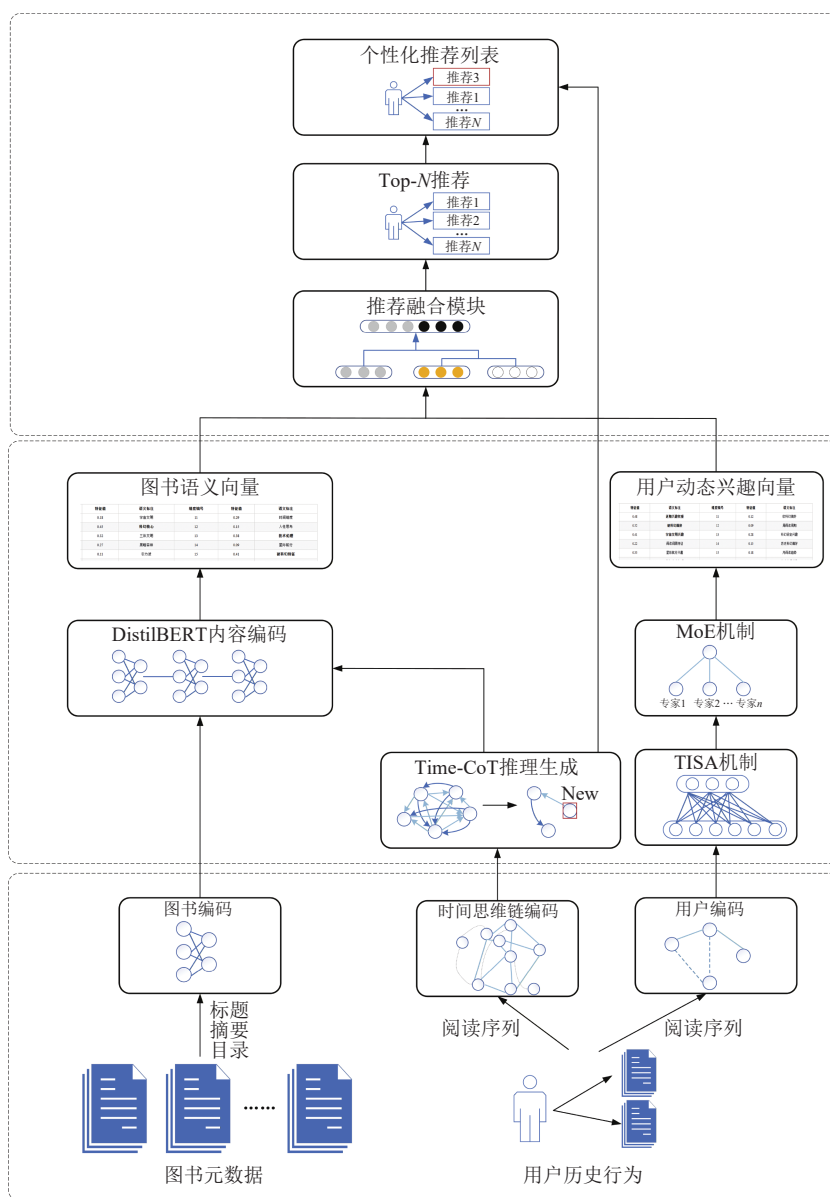


图 1 整体结构图

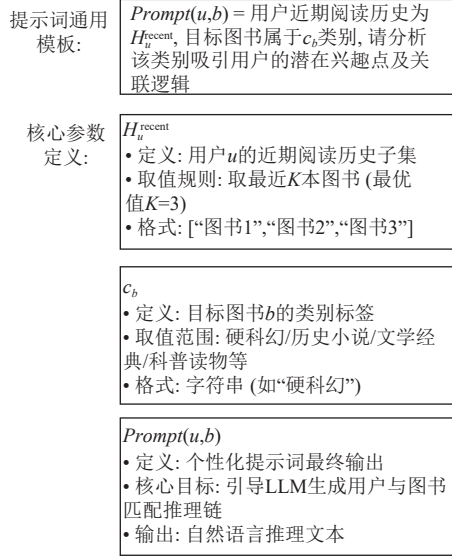


图2 Time-CoT 提示构造逻辑图

生成推理文本后, 需先对其进行标准化预处理, 再通过 DistilBERT 编码为语义向量, 具体步骤如下.

首先对生成的推理文本进行标准化预处理, 通过格式统一 (UTF-8 编码)、冗余信息过滤 (去除重复语句、无意义语气词) 及长度规整 (限制在 512 字符内, 核心推理逻辑优先), 确保输入编码模型的文本一致性, 如式 (6) 所示:

$$CoT_{norm}(u, b) = \text{Normalize}(CoT(u, b)) \quad (6)$$

其中, $\text{Normalize}(\cdot)$ 表示文本标准化操作. 其次采用与图书内容特征提取同源的 DistilBERT-base-uncased 模型, 将标准化后的推理文本编码为语义向量, 通过平均池化层输出 768 维向量, 与图书内容特征向量 $e_b^{content}$ 维度保持一致, 确保编码逻辑同源且无维度冲突, 具体如式 (7) 所示:

$$e_b^{cot} = \text{DistilBERT}_{avg_pool}(CoT_{norm}(u, b)) \quad (7)$$

其中, $\text{DistilBERT}_{avg_pool}(\cdot)$ 表示 DistilBERT 的平均池化编码函数, $e_b^{cot} \in R^{dc}$, 包含图书类别与用户兴趣的关联语义及潜在兴趣点信息.

3.2.3 特征融合

为整合内容特征与思维链特征, 将两者拼接后输入 2 层多层感知机 (multi-layer perceptron, MLP) 进行非线性转换, 生成最终图书语义向量, 如式 (8) 所示:

$$e_b = \text{MLP}(\text{concat}(e_b^{content}, e_b^{cot})) \quad (8)$$

其中, $\text{concat}(\cdot, \cdot)$ 表示向量拼接, $e_b \in R^{dc}$. 该向量不仅

包含图书客观内容 (what), 更融合了用户兴趣关联语义 (why), 为后续推荐的精准性奠定基础.

3.3 时间思维链模块

时间感知思维链模块的功能是基于用户近期阅读历史动态调整思维链的语义聚焦点, 使图书语义向量更贴合用户当前兴趣状态. 模块由 Time-CoT 生成机制、编码增强、创新点分析这 3 部分组成.

3.3.1 Time-CoT 生成机制

用户阅读兴趣随时间推移呈现动态变化, 近期阅读行为对当前兴趣的影响更显著. 设用户 u 的阅读历史序列为 $H_u = \{b_1^{(u)}, b_2^{(u)}, \dots, b_T^{(u)}\}$, $b_t^{(u)}$ 为用户 u 在时间 t 阅读的图书, 则近期阅读历史子集定义如式 (9) 所示:

$$H_u^{recent} = \{b_{T-K+1}^{(u)}, \dots, b_T^{(u)}\} \quad (9)$$

其中, K 为近期历史长度. Time-CoT 生成提示通过参数化构造整合用户近期阅读历史与图书类别, 确保推理链的时间敏感性.

$Prompt(u, b)$ = 用户近期阅读 H_u^{recent} , 目标图书属于 c_b 类别, 请分析该类别吸引用户的潜在兴趣点, 整体如算法 1 所示.

算法 1. Time-CoT 生成机制

输入: 用户阅读历史序列 $H_u = \{b_1^{(u)}, b_2^{(u)}, \dots, b_T^{(u)}\}$, 图书集合 $B = \{b_1, b_2, \dots, b_M\}$, 近期历史长度 K , 图书类别集合 $C = \{c_1, c_2, \dots, c_N\}$, 大语言模型 ChatGLM2-6B, 编码器 DistilBERT, 时间衰减系数 $\alpha=0.1$.

输出: 每本书的 Time-CoT 的推理文本 $CoT(u, b)$, 思维链特征向量集合 $E_{cot} = \{e_b^{cot} | b \in B\}$.

// 获取用户最近 K 本阅读图书, 按阅读时间降序排列

$H_u^{recent} = \text{get_latest_k_books}(H_u, K)$

For each $h \in H_u^{recent}$

// 用 DistilBERT 编码图书文本, 通过 TextRank 提取 Top-5 核心元素

$F_h = \text{TextRank}(\text{DistilBERT.encode}(\text{title}_h \oplus \text{abstract}_h))$

// 计算用户阅读图书 h 至当前的时间间隔 (单位: 天)

$\Delta t_h = t_{\text{current}} - t_h(u)$

// 计算近期阅读图书核心元素交集, 定位用户核心兴趣点

$F_u^{\text{inter}} = \bigcap F_h$

For each target book b in B

// 获取图书 b 的类别标签 (如"硬科幻""历史小说")

$c_b = \text{get_book_category}(b, C)$

$CoT = \text{ChatGLM2-6B}(Prompt(u, b))$

// 调用 ChatGLM2-6B 生成推理文本

$e_b^{cot} = \text{DistilBERT.encode}(CoT(u, b))$

add E_b^{cot} to E_{cot}

End For

End For

Return $CoT(u, b)$ for all (b in B), (E_{cot})

3.3.2 编码增强

通过大语言模型生成 Time-CoT 推理文本后, 采用 DistilBERT 编码为思维链特征向量 e_b^{cot} . 该向量的语义聚焦点随 H_u^{recent} 动态调整: 例如, 若 H_u^{recent} 包含《火星救援》(太空探索类), 则 e_b^{cot} 强调太空技术、极端环境生存等元素. 若 H_u^{recent} 包含《三体》(硬科幻类), 则 e_b^{cot} 强调宇宙文明、黑暗森林法则等元素. 这种动态调整使图书语义向量与用户当前兴趣保持高度关联.

时间思维链模块的主要贡献是实现了时间敏感的语义动态调整. 传统的推荐系统图书的语义是固定的, 不能随着用户的兴趣变化而变化. Time-CoT 模块把用户最近的阅读历史加入思维链的生成中, 使得图书语义向量中除了包含图书本身的内容之外, 还编码了用户当前兴趣与图书的关联逻辑, 提高了推荐的时间相关性以及用户的个性化.

3.4 用户编码器模块

用户兴趣编码模块的目标是捕捉用户兴趣的时间演化特征与兴趣多样性, 并设计时间交互稀疏注意力 (TISA) 与混合专家 (MoE) 双机制结构.

3.4.1 TISA 机制

用户兴趣随时间推移呈现演化规律, 如近期对硬科幻的兴趣高于历史小说, TISA 机制通过时间权重与稀疏注意力捕捉这种演化.

- 时间间隔计算: 设用户 u 阅读图书 $b_t^{(u)}$ 的时间戳为 $t_t^{(u)}$ 当前时间戳为 t_{current} , 则时间间隔定义如式 (10) 所示:

$$\Delta t_t^{(u)} = t_{\text{current}} - t_t^{(u)} \quad (10)$$

其中, $\Delta t_t^{(u)} \geq 0$, 表示用户 u 阅读图书 $b_t^{(u)}$ 至当前的时间差.

- 时间权重计算: 采用指数衰减模型计算历史行为的时间权重, 时间衰减系数 $\alpha=0.1$ (与式 (3) 中 α 含义一致), 如式 (11) 所示:

$$w_t^{(u)} = \frac{\exp(-\alpha \cdot \Delta t_t^{(u)})}{\sum_{t'=1}^T \exp(-\alpha \cdot \Delta t_{t'}^{(u)})} \quad (11)$$

其中, T 为用户 u 的历史行为总数, $w_t^{(u)} \in [0, 1]$ 且满足 $\sum_{t=1}^T w_t^{(u)} = 1$. 时间权重的物理含义为近期行为对当前兴趣的贡献度更高, 如 $\Delta t_t^{(u)} = 7$ 天的行为权重是 $\Delta t_{t'}^{(u)} = 30$

天的 $e^{\alpha \cdot 23}$ 倍.

- 稀疏注意力与时间加权向量: 为降低计算复杂度, 仅保留时间权重 Top- N 的历史行为, 定义稀疏指示函数, 如式 (12) 所示:

$$\delta_t^{(u)} = \begin{cases} 1, & w_t^{(u)} \geq \text{Top-}N \\ 0, & \text{else} \end{cases} \quad (12)$$

将保留的历史行为图书向量按时间权重加权求和, 得到用户时间演化兴趣向量, 如式 (13) 所示:

$$h_u^{\text{TISA}} = \sum_{t=1}^T \delta_t^{(u)} \cdot w_t^{(u)} \cdot e_{b_t^{(u)}} \quad (13)$$

其中, e_b 为图书 $b_t^{(u)}$ 的语义向量, $h_u^{\text{TISA}} \in R^{db}$, 包含用户兴趣随时间演化的特征.

3.4.2 MoE 机制

用户兴趣往往呈现多维度特征, 如同时对硬科幻、历史小说、传记感兴趣, MoE 机制通过多个专家网络分别捕捉不同维度兴趣, 再通过门控网络融合.

- 专家网络设计: 设专家网络数量为 $M=5$, 对应科幻、历史、文学、传记、科普这 5 大类, 每个专家网络 k ($k=1, 2, \dots, M$) 采用 2 层 MLP 结构, 输入为 h_u^{TISA} , 输出为专家兴趣向量, 如式 (14) 所示:

$$h_u^k = \text{MLP}(h_u^{\text{TISA}}) \quad (14)$$

其中, MLP_k 输入维度 $d_b=256$, 隐藏层维度 128, 输出维度 $d_b=256$, ReLU 激活. $h_u^k \in R^{db}$, 表示用户在第 k 个兴趣维度的特征, 如 h_u^1 表示科幻兴趣.

- 门控网络与兴趣融合: 门控网络生成专家网络的权重分配向量, 输入为 h_u^{TISA} , 输出为专家权重向量, 如式 (15) 所示:

$$g_u = \text{Softmax}(\text{MLP}(h_u^{\text{TISA}})) \quad (15)$$

其中, MLP_g 输入维度 $d_b=256$, 隐藏层维度 64, 输出维度 $M=5$, 使用 ReLU 激活函数; $g_u = [g_u^1, g_u^2, \dots, g_u^M]^T$, $g_u^k \in [0, 1]$, 表示用户当前兴趣在各专家维度的分布, 如 $g_u^1 = 0.6$ 表示科幻兴趣占比 60%.

将专家兴趣向量按权重融合, 得到用户最终兴趣向量, 如式 (16) 所示:

$$e_u = \sum_{k=1}^M g_u^k \cdot h_u^k \quad (16)$$

其中, $e_u \in R^{db}$, 包含用户兴趣的时间演化特征与多样性.

TISA 机制解决了传统推荐系统对时间因素建模

的不足,利用时间权重和稀疏注意力捕捉兴趣的时间演化。MoE 机制解决了兴趣多样性建模不足的问题,利用多个专家网络分别捕捉不同的兴趣,再通过门控网络实现多兴趣的动态平衡。双机制相结合使用户向量更加全面地反映当前的兴趣状态,给精准推荐提供可靠的用户表示,整体算法 2 如下。

算法 2. MoE 机制

输入: TISA 编码的用户兴趣随时间演化特征 $h_u^{\text{TISA}} \in R^{256}$, 专家网络数量 $M=5$, 专家网络参数集合 $\theta=\{\theta_1, \theta_2, \dots, \theta_M\}$ (每个专家均采用 2 层 MLP 结构), 门控网络参数集合 ϕ (2 层 MLP), 输入维度 $d_b=256$, 专家 MLP 隐藏层维度 $d_{\text{hidden}}=128$, 门控 MLP 隐藏层维度 $d_{\text{gate}}=64$ 。
输出: 用户动态兴趣向量 $e_u \in R^{d_b}$ 。

```

初始化专家网络  $E_1, E_2, \dots, E_M$  以及参数  $\theta = \{\theta_1, \theta_2, \dots, \theta_M\}$ 
初始化门控网络  $G$  以及参数  $\phi$ 
expert_vectors=[]
// 遍历所有专家网络, 分别提取对应维度兴趣特征
For  $k=1$  to  $M$ 
    // 专家网络  $k$  处理时间演化向量: 第 1 层 MLP 映射+ReLU 激活, 第 2 层 MLP 映射+ReLU 激活, 输出该维度兴趣向量  $h_k^k$ 
     $h_k^k = E_k(h_u^{\text{TISA}}) = \text{ReLU}(W_2^k \cdot \text{ReLU}(W_1^k \cdot h_u^{\text{TISA}} + B_1^k) + B_2^k)$  add  $h_k^k$  to expert_vectors
End For
// 门控网络处理时间演化向量: 两层 MLP 映射+ReLU 激活, 输出原始权重分数  $Z_u$ 
 $Z_u = \text{ReLU}(W_g^2 \cdot \text{ReLU}(W_g^1 \cdot h_u^{\text{TISA}} + B_g^1) + B_g^2)$ 
// 权重归一化: 通过 Softmax 将原始分数转换为概率分布
 $g_u = \text{Softmax}(Z_u)$ 
// 初始化门控权重向量, 与专家网络顺序一一对应
 $\text{gating\_weights} = [g_u^1, g_u^2, \dots, g_u^M]$ 
// 多兴趣融合
 $e_u = \sum \text{gating\_weights}[k] \cdot \text{expert\_vectors}[k]$ 
Return  $e_u$ 

```

3.5 融合推荐模块

融合推荐模块的主要目的是计算用户和图书的语义匹配度,产生个性化的推荐列表,同时输出可以被理解的推荐逻辑,达到精准推荐和透明度并存的效果。

采用余弦相似度度量用户兴趣向量 e_u 与图书语义向量 e_b 的匹配度,相似度分数定义为,如式 (17) 所示:

$$s(u, b) = \cos(e_u, e_b) = \frac{e_u \cdot e_b}{\|e_u\|_2 \cdot \|e_b\|_2} \quad (17)$$

其中, $s(u, b) \in [-1, 1]$, 值越大表示用户 u 与图书 b 的语义匹配度越高。

对图书集合 B 中所有图书计算相似度分数,按降序排序取 Top- N ($N=10$, 实验优化值) 作为推荐列表,如式 (18) 所示:

$$R_u = \text{Top-}N(\{s(u, b)\}) \quad (18)$$

其中, $R_u = \{b_1^*, b_2^*, \dots, b_N^*\}$, b_1^* 为相似度最高的图书,如用户近期读《三体》,则《银河帝国》《星际穿越》等硬科幻图书排在前列。

该模块把可解释性和推荐过程深度结合在一起。传统的推荐系统只输出推荐结果,本模块同时输出推荐逻辑,并且逻辑是依据用户近期的行为和图书语义的动态联系。其价值增强用户信任,推荐逻辑显性化让用户明白推荐理由,减少对系统的怀疑。引导潜在兴趣,推理链中的核心元素,比如宇宙文明,可以引导用户发现潜在的兴趣,拓展阅读范围。支持交互反馈,用户根据自己的喜好调整偏好,比如不喜欢星际政治元素,系统据此对推荐策略进行调整。

4 实验分析

4.1 实验环境

实验环境以图书推荐场景下深层语义挖掘、动态兴趣捕捉以及可解释性生成的实验需求为中心,给模型训练及推理赋予有效稳定的硬件支持与软件匹配,实验环境的配置情况见表 2。

表 2 实验环境配置

硬件配置	参数	软件配置	参数
CPU	Intel Core i7	操作系统	Windows 10 64 位
GPU	NVIDIA GTX 4080Ti×4	开发语言	Python 3.11
硬盘	1 TB	开发工具	PyCharm
内存	512 GB	开发框架	PyTorch 2.0

图书编码器采用轻量化预训练模型 DistilBERT-base-uncased, 含 6 层 Transformer 编码器、768 维隐层维度与 12 个注意力头,通过平均池化输出 768 维内容特征向量,在保证语义捕捉能力的同时提升训练效率。时间思维链编码与内容特征向量维度保持一致,均为 768 维,特征融合层采用 2 层 MLP,输出 256 维最终图书语义向量。

训练时 batch_size 设为 32, 学习率用余弦退火策略动态改变 (开始是 $2E-5$, 最小到 $1E-6$)。每 100 个样本就对模型在验证集上的表现加以检查; 损失函数取对比学习损失和交叉熵损失的加权和 (权重比 0.7:0.3), 加强正负样本的区分。早停策略设为连续 5 轮验证集 NDCG@5 没有提升就结束训练,避免过拟合。

4.2 实验数据

本次实验所采用的图书推荐领域的两个主流公开

数据集为 BookCrossing^[21]和 Goodreads^[22], 分别包含不同规模的用户数、书目种类以及互动情境, 在对数据进行标准化预处理之后, 可以保证实验的全面性与可靠性.

为满足实验需要, 对两类数据集做标准化预处理: 首先, 剔除无有效类别标签、元数据缺失的图书样本以及交互记录少于 3 条的低频用户, 避免数据稀疏性对模型训练造成影响. 其次对图书元数据中的标题、摘要、目录文本进行拼接, 统一编码格式并去除无意义符号和重复内容. 最后, 保留用户交互记录中时间戳信息, 按阅读时间先后排序构建用户历史行为序列. 预处理之后, BookCrossing 数据集中留下有效的交互记录 4 万条. Goodreads 数据集保留有效的用户交互记录 6 万条.

两组数据集用 8:1:1 的比例进行随机划分, 分别为训练集、验证集和测试集, 在训练集中做模型的参数优化, 在验证集中调参、模型选择, 最后用测试集做最终的性能评价.

4.3 评价指标

为了从推荐精准度、排序合理性、可解释度这 3 个方面来全方位评价 Time-CoT-LLM 模型在推荐上的表现, 根据图书推荐任务的特性以及模型的主要目标, 选取了推荐效果指标和可解释性指标这两大类评价标准, 并对每个标准进行了具体的定义和计算方式的说明.

推荐效果指标聚焦模型的精准推荐能力与排序质量, 采用推荐系统领域主流的 AUC 、 $NDCG@K$ 、 $Recall@K$ 作为核心评价指标, 覆盖二分类预测准确性与 Top- N 推荐合理性.

AUC 用于衡量模型区分用户感兴趣图书与不感兴趣图书的能力, 通过计算 ROC 曲线下面积量化模型的二分类预测性能. 设测试集中用户与图书交互对总数为 N , 其中正例 (用户感兴趣) 为 P 个, 负例 (用户不感兴趣) 为 $N-P$ 个, 定义排序得分 $s(u, b)$ 为用户 u 与图书 b 的语义匹配度, 则 AUC 计算如式 (19) 所示:

$$AUC = \frac{1}{p \cdot (N - P)} \sum \sum I(s(u, b_+) > s(u, b_-)) \quad (19)$$

其中, $I(\cdot)$ 为指示函数, 当括号内条件成立时取值 1, 否则取值 0. AUC 取值范围为 [0.5, 1], 值越接近 1 表示模型区分能力越强.

$NDCG@K$ 衡量 Top- N 推荐列表的排序合理性, 考虑用户对图书的兴趣程度权重, 通过折扣累积增益的归一化处理消除推荐列表长度影响. 设 K 为推荐列表

长度 (实验取 $K=5$). 兴趣得分 (评分数据中取原始评分, 无评分交互取 1, 非交互取 0), 则 $DCG@K$ 与 $NDCG@K$ 计算如式 (20)、式 (21) 所示:

$$DCG@K = \sum_{i=1}^K \frac{2^{r_{ui}-1}}{\log_2(i+1)} \quad (20)$$

$$NDCG@K = \frac{DCG@K}{IDCG@K} \quad (21)$$

其中, $IDCG@K$ 为理想 $DCG@K$, 即按用户真实兴趣得分降序排列的 Top- N 列表对应的 DCG 值. $NDCG@K$ 取值范围为 [0, 1], 值越大表示 Top- N 推荐的排序越贴合用户真实偏好.

$Recall@K$ 衡量 Top- N 推荐列表中覆盖用户真实感兴趣图书的比例, 反映模型的召回能力. 设测试集中用户 u 真实感兴趣的图书集合为 G_u , 推荐列表中属于 G_u 的图书数量为 $|G_u \cap R_{u,K}|$, 则 $Recall@K$ 计算如式 (22) 所示:

$$Recall@K = \frac{1}{|U|} \sum_{u \in U} \frac{|G_u \cap R_{u,K}|}{|G_u|} \quad (22)$$

其中, U 为测试集用户集合, $R_{u,K}$ 为用户 u 的 Top- N 推荐列表. $Recall@K$ 取值范围为 [0, 1], 值越大表示模型越能精准召回用户潜在感兴趣的图书.

理由相关性衡量时间思维链 (Time-CoT) 与用户行为、图书特征的关联程度, 避免推荐理由空洞化. 设思维链文本 $CoT(u, b)$ 中包含的用户近期行为关联词数量为 C_1 , 如用户近期阅读图书的标题、核心元素, 图书特征关联词数量为 C_2 如图书类别、主题、核心元素, 思维链总词数为 L , 则 RR 计算如式 (23) 所示:

$$RR = \frac{C_1 + C_2}{L} \quad (23)$$

其中, 关联词通过 TextRank 提取核心关键词, 结合词频统计过滤后自动匹配, 基于图书元数据与用户历史行为文本的词频统计自动识别, RR 取值范围为 [0, 1], 值越大表示推荐理由与用户行为、图书特征的关联性越强.

用户满意度通过人工评测获取, 邀请 100 名测试集用户对推荐列表的理由清晰度、可信度进行 5 分制评分 (1 分=完全不认可, 5 分=完全认可), 计算所有用户评分的平均值, 如式 (24) 所示:

$$US = \frac{1}{100} \sum_{i=1}^{100} score_i \quad (24)$$

其中, $score_i$ 为第 i 名用户的评分. US 取值范围为 $[1, 5]$, 值越大表示用户对推荐理由的认可度与信任度越高.

4.4 实验分析

4.4.1 超参数调优实验

超参数调优实验聚焦 $batch_size$ 与 $epoch$ 的组合优化, 以验证集 $NDCG@5$ 为核心评价指标, 兼顾训练稳定性与泛化能力, 在 BookCrossing 数据集上测试不同参数组合的模型性能, 筛选适配 Time-CoT-LLM 架构的最优训练配置, 实验结果如图 3 所示.

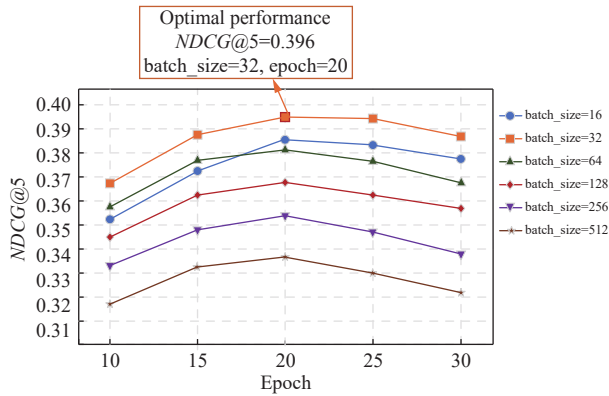


图 3 不同 $batch_size$ 与 $epoch$ 组合在验证集上的 $NDCG@5$ 模型性能

实验结果表明, 当 $batch_size$ 为 32、 $epoch$ 为 20 时, 模型的性能最好, $NDCG@5$ 为 0.396, 是特征学习充分性和泛化能力最好的组合. 从 $batch_size$ 的角度来看, 在 16–32 区间内, 模型性能呈上升趋势, 32 以后随着 $batch_size$ 的增大持续下滑, 尤其是 256、512 等超大批次, 由于单次迭代样本同质化严重, 不能很好地捕捉到用户兴趣的时间演变和细粒度语义联系, 梯度更新没有针对性. 从 $epoch$ 维度看, 所有 $batch_size$ 在 20 轮达到性能峰值后由于训练数据噪声拟合造成性能下降, 并且 $batch_size$ 越小过拟合越严重, 超大 $batch_size$ 虽然过拟合程度小, 但是整体性能上限低.

为了验证时间衰减系数 α 对模型性能的影响, 进一步说明 $\alpha=0.1$ 的合理性, 本文在 BookCrossing 数据集上设计敏感性分析实验: 固定 $batch_size=32$ 、 $epoch=20$, 设置 α 取值为 0.02、0.05、0.1、0.15、0.2、0.3, 对比不同 α 下模型的核心推荐指标 (AUC 、 $NDCG@5$ 、 $Recall@5$) 与可解释性指标 (RR 、 US), 实验结果如表 3 所示.

实验结果清楚地表现出 α 对于模型性能的非线性影响, 证明了图书阅读场景下时间衰减系数的适配逻

辑. α 太小 (≤ 0.05), 时间衰减效应微弱, 远期阅读行为权重占比过高, 模型很难抓住用户当下的核心兴趣, 推荐精准度与理由相关性都较低. α 过大的时候 (≥ 0.2), 衰减效应过大, 模型过分依赖最近的行为, 忽略用户的长期稳定兴趣, 造成推荐的多样性下降, 指标明显下降. 当 $\alpha=0.1$ 的时候, 模型的精确度和可解释性达到最佳状态, 突出了近期阅读行为对当前兴趣的影响, 又保留了历史行为的参考价值, 完全符合图书阅读中用户兴趣渐变而不是突变的特点.

表 3 时间衰减系数实验

α 取值	AUC	$NDCG@5$	$Recall@5$	RR	US
0.02	0.743	0.361	0.263	0.78	3.9
0.05	0.761	0.382	0.275	0.80	4.1
0.1	0.772	0.396	0.285	0.82	4.3
0.15	0.768	0.391	0.281	0.81	4.2
0.2	0.755	0.378	0.272	0.79	4.0
0.3	0.739	0.359	0.258	0.76	3.8

4.4.2 消融实验

消融实验旨在验证 Time-CoT-LLM 模型各核心模块, 时间思维链模块 Time-CoT、时间稀疏注意力 TISA、混合专家 MoE、LLM 增强图书编码的有效性, 通过逐步移除各模块构建变体模型, 在 BookCrossing 和 Goodreads 数据集上对比核心推荐指标与可解释性指标, 实验结果如表 4 所示.

完整的模型在两类数据集的所有指标上都表现最好, 说明各个核心模块有协同作用. 移除 Time-CoT 模块后, AUC 、 $NDCG@5$ 分别下降 3.9%–4.2%、9.6%–10.2%, 可解释性指标 (RR 、 US) 降幅最大 (超过 45%), 说明时间思维链不但是推荐理由的生成载体, 也是通过动态语义关联提高了用户和图书的匹配准确度, 它对可解释性的贡献远大于其他模块. 移除 TISA 或者 MoE 机制之后, 模型的性能都会有不同程度的下降, TISA 对于排序指标 $NDCG@5$ 的影响更大, MoE 对于召回率 $Recall@5$ 的提升更加明显, 证明 TISA 对于时间的捕捉以及 MoE 对于兴趣多样性的建模缺一不可. 移除 LLM 增强之后, 图书语义编码的精度下降, 使得推荐的匹配度降低, RR 指标由于缺少深层语义关联而下降, 证明了 LLM 对于挖掘图书细粒度特征的不可替代性. 各个模块的消融结果表明, Time-CoT-LLM 的主要优势是由于 4 个模块的相互配合而产生的, 准确地解决了动态兴趣捕捉、深层语义理解和可解释性这 3 个主要问题.

表4 消融实验

模型变体	数据集	AUC	NDCG@5	Recall@5	RR	US
完整模型 (Time-CoT-LLM)	BookCrossing	0.772	0.396	0.285	0.82	4.3
	Goodreads	0.785	0.403	0.292	0.84	4.4
移除Time-CoT模块	BookCrossing	0.741	0.358	0.261	0.41	3.1
	Goodreads	0.753	0.365	0.268	0.43	3.2
移除TISA机制 (保留MoE)	BookCrossing	0.755	0.372	0.270	0.81	4.0
	Goodreads	0.768	0.379	0.277	0.83	4.1
移除MoE机制 (保留TISA)	BookCrossing	0.748	0.364	0.259	0.80	3.9
	Goodreads	0.761	0.371	0.266	0.82	4.0
移除LLM增强 (用传统 TF-IDF编码图书)	BookCrossing	0.732	0.349	0.248	0.58	3.5
	Goodreads	0.745	0.356	0.255	0.60	3.6

4.4.3 对比实验

对比实验选取 5 类主流推荐模型作为基线, 在 BookCrossing 和 Goodreads 数据集上进行性能比拼, 核心验证 Time-CoT-LLM 在推荐精准性与可解释性上的优势, 实验结果如表 5 所示。

本文提出的 Time-CoT-LLM 在两个数据集上所有指标都是最好的, 比传统的协同过滤 AUC 提高了 4.1%–6.9%, NDCG@5 提高了 16.8%–19.7%, 优势很明显。传统协同过滤由于没有考虑语义和时间的建

模, 所以性能垫底且无解释性。BERT 所基于的模型可以捕捉到一部分语义, 但是缺乏时间动态和可解释性。纯 GPT-3.5 模型语义理解能力较强, 但是推荐理由和用户行为无关。TALLRec 通过 LLM 调优提高适配性, 但是没有解决动态兴趣和细粒度语义问题。Time-CoT-LLM 依靠 4 大模块协同工作, 除了 TISA 和 MoE 准确捕捉用户兴趣外, 更用 Time-CoT 把行为和语义关联起来, 从而实现了精准性、可解释性的双重突破。

表5 对比实验

模型变体	数据集	AUC	NDCG@5	Recall@5	RR	US
协同过滤 (CF) ^[23]	BookCrossing	0.703	0.332	0.228	—	—
	Goodreads	0.715	0.339	0.235	—	—
BERT-based序列推荐 ^[24]	BookCrossing	0.738	0.365	0.254	0.45	3.2
	Goodreads	0.751	0.372	0.261	0.47	3.3
纯LLM 模型 (GPT-3.5) ^[25]	BookCrossing	0.741	0.368	0.257	0.52	3.4
	Goodreads	0.754	0.375	0.264	0.54	3.5
TALLRec (LLM 调优框架) ^[26]	BookCrossing	0.753	0.379	0.266	0.61	3.7
	Goodreads	0.766	0.386	0.273	0.63	3.8
Time-CoT-LLM (本文模型)	BookCrossing	0.772	0.396	0.285	0.82	4.3
	Goodreads	0.785	0.403	0.292	0.84	4.4

4.4.4 推荐覆盖实验

推荐覆盖实验聚焦模型在不同图书领域的适配能力, 选取科幻、历史、文学、传记、科普这 5 大核心类别, 在 Goodreads 数据集上测试 Time-CoT-LLM 与最优基线模型 TALLRec 的性能差异, 验证模型对多领域图书的推荐覆盖效果, 实验结果如表 6 所示。

Time-CoT-LLM 在 5 个图书领域都具有较好的稳定性和优异的性能, 相较于 TALLRec 各指标平均提高 3.2%–4.5%。其中文学和科幻领域提升最明显, 因为文学和科幻领域的图书细粒度差异较大, Time-CoT-LLM 依靠 LLM 增强编码器精准捕捉深层语义, 并且利用 MoE 机制适配多维度兴趣。即便是在传记、科普等小众的领域, 模型依然有着高效的推荐能力, 证明了

其并不局限于热门的领域, 可以有效地覆盖各种各样的图书。该结果证明了模型对各个领域图书的通用性。

表6 推荐覆盖实验

图书领域	模型	AUC	NDCG@5	Recall@5
科幻 (含硬/软科幻)	TALLRec	0.762	0.381	0.269
	Time-CoT-LLM	0.789	0.408	0.297
历史	TALLRec	0.758	0.376	0.265
	Time-CoT-LLM	0.783	0.401	0.291
文学	TALLRec	0.765	0.383	0.271
	Time-CoT-LLM	0.790	0.410	0.299
传记	TALLRec	0.751	0.370	0.260
	Time-CoT-LLM	0.778	0.395	0.286
科普	TALLRec	0.749	0.368	0.258
	Time-CoT-LLM	0.775	0.392	0.283

4.4.5 推荐效率实验

推荐效率实验聚焦模型在实际应用中的部署可行

性,以参数量(M)和单用户推理时间(ms)为核心指标,在BookCrossing数据集上对比Time-CoT-LLM与基线模型的效率差异,输入为用户最近20条历史行为,输出Top-10推荐列表,结果如图4所示。

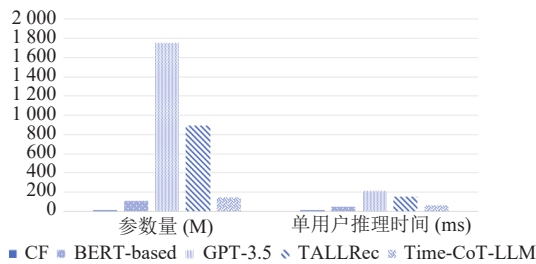


图4 推荐效率实验

传统协同过滤参数最少、推理速度最快,但是性能较差;纯LLM模型和TALLRec由于参数量太大(分别为1750M和890.3M),推理时间远远超过实用阈值。Time-CoT-LLM采用DistilBERT轻量化设计及稀疏注意力机制,把参数量控制在145.8M,比纯LLM模型少8.3%,推理时间62.8ms,满足实时推荐的需求(≤ 100 ms)。效率优势来自模块协同优化,TISA机制削减历史行为计算量,MoE根据需要开启专家网络,并非全部参数都参与推理。

4.4.6 可解释性实验

可解释性实验聚焦模型推荐理由的合理性与用户信任度,选取理由相关性(RR)与用户满意度(US)为核

心指标,在Goodreads数据集上对比Time-CoT-LLM与各基线模型的可解释性表现,实验邀请100名用户参与满意度评分,结果如表7所示。

表7 各模型可解释性指标对比

模型	RR	US
传统协同过滤(CF)	—	—
BERT-based序列推荐	0.45	3.2
纯LLM模型(GPT-3.5)	0.52	3.4
TALLRec(LLM调优框架)	0.61	3.7
Time-CoT-LLM(本文模型)	0.84	4.4

传统协同过滤没有给出推荐的理由,即不可解释。BERT基于的模型由于没有关联用户行为,所以理由相关性低。纯LLM模型可以生成自然语言的理由,但是缺少了行为的关联,导致用户信任度不高。TALLRec利用LLM调优提高了语义关联性,但是推理链静态化影响了满意度。Time-CoT-LLM用动态生成的时间思维链将用户近期的行为和图书的特征联系起来,RR值为0.84,US为4.4,85%的用户认为它理由清晰。这证明了时间思维链机制是有效的,使推荐理由由泛化的描述变成了个性化的推理,更加符合用户的信任感。

4.4.7 具体案例分析

为了直观地证明Time-CoT生成的推荐理由同用户的行为之间存在联系,选择不同兴趣偏好的测试集用户,展示最近一段时间内的阅读行为、Time-CoT生成的内容以及推荐的结果,具体的案例如表8所示。

表8 对比实验

用户ID	近期阅读行为(含时间戳)	目标推荐图书	Time-CoT生成内容	理由相关性(RR)	用户满意度评分(5分制)
U001	2024-05-10《三体》(硬科幻) 2024-05-18《球状闪电》(硬科幻) 2024-05-25《流浪地球》(硬科幻)	《银河帝国》(硬科幻)	用户近期3本阅读均为硬科幻类,核心元素交集为“宇宙文明”“科幻设定”,对《三体》的兴趣权重达42%。目标图书属于硬科幻类别,包含星际政治、文明兴衰等核心元素,与用户关注的宏大宇宙叙事主题高度契合,符合用户对硬科幻设定的持续兴趣。	0.89	5
U002	2024-04-20《明朝那些事儿》(历史) 2024-04-28《万历十五年》(历史) 2024-05-05《显微镜下的大明》(历史)	《中国历代政治得失》(历史)	用户近期阅读聚焦历史类图书,核心元素交集为“古代政治”“历史细节”,对《万历十五年》的兴趣权重为38%。目标图书属于历史类别,围绕历代政治制度演变展开,深度剖析制度设计与社会发展的关联,与用户关注的历史细节探索需求匹配,延续用户对历史题材的探索兴趣。	0.85	4
U003	2024-05-02《人类简史》(科普) 2024-05-09《时间简史》(科普) 2024-05-16《万物简史》(科普)	《宇宙简史》(科普)	用户近期阅读均为科普类图书,核心元素交集为“自然规律”“宇宙探索”,对《时间简史》的兴趣权重达45%。目标图书属于科普类别,以通俗语言解读宇宙起源与演化规律,涵盖黑洞、时空弯曲等核心知识点,与用户热衷的科学探索主题高度相关,契合用户对科普内容的深度阅读需求。	0.87	5

上述案例很好地体现出 Time-CoT 生成的内容和用户行为之间有很强的联系。Time-CoT 并不是泛化的图书类别描述,而是一种精确地提取用户最近的阅读核心元素交集和时间权重的方法,从而保证推荐的理由是根据用户的真实行为来提出的。U001 的推荐理由中提到“硬科幻”“宇宙文明”等高频关联词,并且明确提到了《三体》这一部高权重的影响。U002 的推理集中于“古代政治”“历史细节”等用户一直关注的话题。基于用户行为的个性化推理,可以解释推荐的理由是客观存在的,同时又加强了用户对于推荐结果的信任感,使 Time-CoT 在提高推荐可解释性方面具有明显的优势。

5 结论与展望

本文针对图书推荐中动态兴趣捕捉弱、深层语义理解不足和可解释性差这 3 个主要问题,提出 Time-CoT-LLM 模型。该模型把 LLM 加强的图书编码、时间思维链、TISA 与 MoE 联合的用户编码、融合推荐这 4 大模块融合起来,在两个公开的数据集上对 *AUC*、*NDCG@5* 等指标进行了验证,并且明显提高了推荐的可解释性,给文化领域的智能推荐提供了一种有效的解决办法。

未来可以从以下 3 个方面来改进: (1) 增加多模态图书特征,加入封面、作者访谈等信息提高语义捕获准确度。(2) 优化冷启动策略,依靠少量用户行为和图书元数据迅速生成个性化推荐。(3) 加入用户反馈,根据用户的反馈来动态调整思维链的生成逻辑和兴趣建模权重,从而进一步提高模型的自适应能力及实用性。

参考文献

- 1 Cho J, Hyun D, Kang S, *et al.* Learning heterogeneous temporal patterns of user preference for timely recommendation. *Proceedings of the 2021 Web Conference*. Ljubljana: ACM, 2021. 1274–1283.
- 2 Huang X, Hou HY, Sun MY. A novel temporal recommendation method based on user query topic evolution. *Knowledge-based Systems*, 2022, 241: 108239. [doi: [10.1016/j.knsys.2022.108239](https://doi.org/10.1016/j.knsys.2022.108239)]
- 3 Ahmed I, Ahmad M, Chehri A, *et al.* A heterogeneous network embedded medicine recommendation system based on LSTM. *Future Generation Computer Systems*, 2023, 149: 1–11.
- 4 Qiu ZP, Wu X, Gao JY, *et al.* U-BERT: Pre-training user representations for improved recommendation. *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. AAAI Press, 2021. 4320–4327.
- 5 Zhang JJ, Xie RB, Hou YP, *et al.* Recommendation as instruction following: A large language model empowered recommendation approach. *ACM Transactions on Information Systems*, 2025, 43(5): 114. [doi: [10.1145/3795143](https://doi.org/10.1145/3795143)]
- 6 Petruzzelli A. Towards symbiotic recommendations: Leveraging LLMs for conversational recommendation systems. *Proceedings of the 18th ACM Conference on Recommender Systems*. Bari: ACM, 2024. 1361–1367.
- 7 Liu QJ, Chen N, Sakai T, *et al.* ONCE: Boosting content-based recommendation with both open- and closed-source large language models. *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. Merida: ACM, 2024. 452–461.
- 8 Yuan W, Yang CQ, Ye GH, *et al.* FELLAS: Enhancing federated sequential recommendation with LLM as external services. *ACM Transactions on Information Systems*, 2025, 43(6): 144. [doi: [10.1145/3709138](https://doi.org/10.1145/3709138)]
- 9 Lin XY, Wang WJ, Li YQ, *et al.* Data-efficient fine-tuning for LLM-based recommendation. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Washington: ACM, 2024. 365–374.
- 10 Abu-Rasheed H, Abdulsalam MH, Weber C, *et al.* Supporting student decisions on learning recommendations: An LLM-based chatbot with knowledge graph contextualization for conversational explainability and mentoring. *arXiv:2401.08517*, 2024.
- 11 王大阜, 邓志文, 贾志勇, 等. 基于用户画像的高校图书馆个性化图书推荐研究. *河南师范大学学报(自然科学版)*, 2022, 50(3): 95–103.
- 12 Ahmed E, Letta A. Book recommendation using collaborative filtering algorithm. *Applied Computational Intelligence and Soft Computing*, 2023, 2023: 1514801.
- 13 易晓宇, 易绵竹. 基于兴趣信息深度融合的网络图书资源推荐. *吉林大学学报(工学版)*, 2024, 54(12): 3614–3619. [doi: [10.13229/j.cnki.jdxbgxb.20231025](https://doi.org/10.13229/j.cnki.jdxbgxb.20231025)]
- 14 王莉军, 王淑君, 高影繁. 一种基于图结构的相似图书内容推荐方法. *情报科学*, 2025, 43(5): 199–205. [doi: [10.13833/j.issn.1007-7634.2025.05.020](https://doi.org/10.13833/j.issn.1007-7634.2025.05.020)]
- 15 Devika P, Milton A. Book recommendation using sentiment analysis and ensembling hybrid deep learning models. *Knowledge and Information Systems*, 2025, 67(2):

- 1131–1168. [doi: [10.1007/s10115-024-02250-z](https://doi.org/10.1007/s10115-024-02250-z)]
- 16 Lyu HJ, Jiang S, Zeng HQ, *et al.* LLM-Rec: Personalized recommendation via prompting large language models. Proceedings of the 2024 Findings of the Association for Computational Linguistics. Mexico City: ACL, 2024. 583–612.
- 17 Lin JY, Peng S, Zhang Z, *et al.* TLRec: A transfer learning framework to enhance large language models for sequential recommendation tasks. Proceedings of the 18th ACM Conference on Recommender Systems. Bari: ACM, 2024. 1119–1124.
- 18 Ji JC, Li ZL, Xu SY, *et al.* GenRec: Large language model for generative recommendation. Proceedings of the 46th European Conference on Information Retrieval Advances in Information Retrieval. Glasgow: Springer, 2024. 494–502.
- 19 龚镇辉, 史晓雨, 鲁云, 等. 面向敏感属性缺失的大语言模型公平推荐框架. 计算机应用. <https://link.cnki.net/urlid/51.1307.TP.20251121.1217.002>. (2025-11-24).
- 20 于恩海, 温彦, 陈宇翱. 基于大语言模型与图神经网络的会话推荐增强框架. 计算机应用研究, 2026, 43(1): 35–42.
- 21 Luo SC, Yao YX, He BW, *et al.* Integrating large language models into recommendation via mutual augmentation and adaptive aggregation. IEEE Journal of Selected Topics in Signal Processing, 2026, 20(1): 77–89. [doi: [10.1109/JSTSP.2026.3653160](https://doi.org/10.1109/JSTSP.2026.3653160)]
- 22 Acharya A, Singh B, Onoe N. LLM based generation of item-description for recommendation system. Proceedings of the 17th ACM Conference on Recommender Systems. Singapore: ACM, 2023. 1204–1207.
- 23 Anwar T, Uma V, Hussain MI, *et al.* Collaborative filtering and kNN based recommendation to overcome cold start and sparsity issues: A comparative analysis. Multimedia Tools and Applications, 2022, 81(25): 35693–35711. [doi: [10.1007/s11042-021-11883-z](https://doi.org/10.1007/s11042-021-11883-z)]
- 24 Noorian A. A BERT-based sequential POI recommender system in social media. Computer Standards & Interfaces, 2024, 87: 103766. [doi: [10.1016/j.csi.2023.103766](https://doi.org/10.1016/j.csi.2023.103766)]
- 25 Wang SF. Intelligent recommendation of open education teaching resources based on hybrid collaborative recommendation algorithm with large language model. Proceedings of the 2025 International Conference on Intelligent Systems and Computational Networks. Bidar: IEEE, 2025. 1–9.
- 26 Bao KQ, Zhang JZ, Zhang Y, *et al.* TALLRec: An effective and efficient tuning framework to align large language model with recommendation. Proceedings of the 17th ACM Conference on Recommender Systems. Singapore: ACM, 2023. 1007–1014.
- (校对责编: 李慧鑫)