未来金融创新工程中心
Center for Future Financial Innovation

CFFI 工作论文 No. 26-01

中国A股上市公司行业分类数据集构建
——基于大语言模型的方法

吴　轲
中国人民大学未来金融创新工程中心
中国人民大学财政金融学院

应镇焜
中国人民大学财政金融学院

钱宗鑫
中国人民大学未来金融创新工程中心
中国人民大学财政金融学院

周德馨
纽约市立大学巴鲁克学院
齐克林商学院

中国人民大学未来金融创新工程中心

本论文可从中国人民大学未来金融创
新工程中心电子论文库免费下载：
http://cffi.ruc.edu.cn/kycg/gzlw

中国 A 股上市公司行业分类数据集构建
——基于大语言模型的方法

Construction of an Industry Classification Dataset for Chinese A-Share Listed Companies: A Large Language Model Approach

吴　轲[1]　　应镇焜[2]　　钱宗鑫[3]　　周德馨[4]

[1] 吴轲，中国人民大学未来金融创新工程中心，中国人民大学财政金融学院，电子邮件：ke.wu@ruc.edu.cn。

[2] 应镇焜，中国人民大学财政金融学院，电子邮件：yingzhenkun0205@ruc.edu.cn。

[3] 钱宗鑫（通讯作者），中国人民大学未来金融创新工程中心，中国人民大学财政金融学院，电子邮件：qianzx@ruc.edu.cn。

[4] 周德馨，纽约市立大学巴鲁克学院，齐克林商学院，电子邮件：dexin.zhou@baruch.cuny.edu。

中国 A 股上市公司行业分类数据集构建
——基于大语言模型的方法

摘要：行业分类是金融经济学实证研究的基础性工具，但现有中国 A 股市场多套行业分类标准普遍存在更新滞后、区分度不足等问题。本文基于 2007 至 2023 年 52702 份 A 股上市公司年报“管理层讨论与分析”（MD&A）文本，利用大语言模型的文本嵌入能力与层次聚合聚类算法，构建了一套涵盖 26 个一级、102 个二级和 271 个三级行业的中国 A 股上市公司行业分类数据集。实证结果显示，该分类体系在行业间差异性和行业内相似性两个维度上均显著优于中上协、申万和万得等主流分类标准，表明其能够更有效地实现“类内相似、类间差异”的分类目标。拓展性分析表明，基于 LLM 分类构造的领先–滞后对冲投资组合能够产生统计显著的月度平均收益，并且在 Fama–French 五因子和中国四因子模型调整后仍然显著；Fama–MacBeth 回归进一步证实，LLM 分类在捕获高价股同行业动量效应方面具有最强的预测能力，为该分类体系的准确性提供了基于资产定价的证据。本文为中国 A 股市场提供了一套分类精准、数据驱动、可动态更新的行业分类框架，为公司金融、资产定价等实证研究提供新的分析工具。

关键词：上市公司　行业分类　大语言模型　文本嵌入

一、引言

行业分类是金融经济学与实证会计研究的基石性工具。行业效应能够解释公司利润差异中相当可观的部分（McGahan and Porter，1997），在面板回归中控制行业固定效应是公司金融与资产定价领域最为常见的研究设计。然而，现有中国 A 股市场的行业分类体系普遍存在更新滞后和区分度不足等问题，已逐渐难以满足当前对精准行业分类及动态产业趋势研判的需求。

目前，中国 A 股市场并存着中国上市公司协会（中上协，原证监会）分类、申万分类、万得分类等多套行业分类标准，各标准在分类逻辑、层级结构和更新频率上差异显著。中上协分类作为官方管理型标准，其框架直接参照《国民经济行业分类》国家标准，首要目标是服务于国民经济统计和行政监管，而非金融研究和投资分析，其更新机制相对滞后，难以及时反映新兴产业的快速涌现。申万、万得等投资型分类标准同样也有更新滞后的问题，尽管在市场适应性方面有所改善，但其编制过程高度依赖人工判断，缺乏公开透明的分类算法。

基于上述事实，中国学术界与市场目前尚缺乏一套分类准确，基于公开数据和方法构建，并能够高频动态更新、捕捉新兴业态的行业分类体系。Hoberg 和 Phillips（2016）开创性地利用美国上市公司年报文本构建了动态 TNIC 行业分类，实证表明其比传统的 SIC 和 NAICS 分类能更准确地刻画企业间的竞争关系与行业边界。这种基于上市公司公开文本的行业分类体系构建为中国市场的新型行业分类提供了重要的方法论借鉴。同时，近年来大语言模型技术的飞速发展，为处理上市公司文本、实现高精度的行业分类提供了技术工具。因此，本研究旨在利用中国上市公司年报这一公开文本数据，使用大语言模型构建一套分类准确、可高频动态更新、方法透明可复现的行业分类体系，以填补这一空白。

本文基于 2007 至 2023 年中国 A 股上市公司年报中“管理层讨论与分析”（MD&A）章节的 52702 份文本，利用大语言模型的文本嵌入能力对各公司 MD&A 文本进行语义向量化，并通过层次聚合聚类方法构建三级行业分类体系。在得到分类结果后，本文借助大语言模型设计了“局部摘要-全局命名”的两阶段提示词策略对行业进行语义命名，最终形成了一套涵盖 26 个一级、102 个二级和 271 个三级行业的中国 A 股上市公司行业分类数据集。

本文的实证分析从多个维度验证了该分类体系的准确性。首先，在行业间差异性分析中，本文发现在同等颗粒度下，LLM 分类体系在营业利润率、资产回报率、营业收入增长率和资本支出增长率等财务指标上的行业间标准差均显著高于申万、万得和中上协等主流分类标

准，表明该方法能够更有效地将财务特征相异的公司归入不同行业类别。其次，在行业内相似性分析中，行业哑变量回归的 R^2 显示 LLM 分类体系在多数指标上具有更高的解释力，能够更好地将财务特征相近的公司汇聚于同一行业。上述两组分析共同表明，本文构建的分类体系在“类内相似、类间差异”这一分类质量的核心标准上具有显著优势。

接下来，在拓展性分析中，本文进一步检验了该分类体系在实证资产定价框架下的有效性。结果显示，基于 LLM 二级和三级分类构造的领先-滞后对冲投资组合能够产生月均 1.29% 和 1.53%的等权平均收益，且在 Fama-French 五因子和中国四因子模型调整后仍然显著，而其余分类体系在同等检验框架下均未能超越基于 LLM 的分类体系。基于流通市值加权的投资组合也呈现出类似结果。在 Fama-MacBeth 横截面回归中，LLM 二级分类在捕获高价股同行业动量效应方面具有最强的预测能力，交乘项系数在控制了资产增长率、公司规模、账面市值比和毛利率后仍显著，而其他分类体系的交乘项系数均不显著。

本文相比现有文献主要有以下三点贡献。首先，本文构建了一套涵盖三级行业分类的中国 A 股上市公司行业分类数据集，填补了中国市场缺乏方法透明、公开数据驱动、可动态更新的行业分类体系这一研究空白。其次，研究方法上，本文通过行业间差异性、行业内相似性、投资组合构造和 Fama-MacBeth 回归等多维度实证检验，系统地将本文分类体系与申万、万得和中上协等主流标准进行了定量比较，为中国市场行业分类质量的评估提供了可复制的方法框架。最后，本文构建的分类体系为实证研究与金融实务提供了更具区分度的行业测度，并在方法论上突破了传统行业分类的规则驱动范式，开创了大模型结合的数据驱动方法，进一步拓展了大语言模型在金融文本分析中的应用边界。

本文结构安排如下：第二部分梳理现有行业分类标准及其局限；第三部分为文献综述；第四部分介绍数据来源与分类体系构建流程；第五部分为实证检验及分析；第六部分为拓展性分析；第七部分为本文的研究结论。

二、现有行业分类标准

目前，中国 A 股市场并存着多种行业分类标准，按其制定主体与目的可大致分为两类：一是基于国民经济统计的官方或半官方标准，如中国上市公司协会（中上协）行业分类；二是服务于投资分析与指数编制的投资型标准，如申万行业分类、万得（Wind）行业分类等。此外，中信行业分类、中证指数分类、恒生指数分类以及国际通行的全球行业分类系统（GICS）也在特定领域被广泛使用。尽管这些标准在各自的应用场景中发挥了重要作用，但它们普遍

采用基于主营业务的静态划分方法，在应对市场快速变化、产业结构升级和企业跨界转型时，逐渐暴露出更新滞后、结构固化和区分度不足的缺点。本节将对主流标准进行简要梳理，并系统阐述其固有局限，为后续构建基于大语言模型的动态分类体系提供问题背景。

（一）中国上市公司协会（中上协）行业分类

中国上市公司协会于 2023 年 5 月施行了《中国上市公司协会上市公司行业统计分类指引》，该文件在遵循 2012 年证监会标准框架的基础上，将分类原则、更新周期等具体细节进行了进一步明确和细化，成为当前执行的官方分类标准。

在分类结构上，中上协严格参照《国民经济行业分类》国家标准（GB/T 4754-2017），构建起门类、大类和中类三个层级。其中，门类以一位拉丁字母表示（如 C 代表制造业），大类以两位数字表示，中类则以三位数字表示。目前，该分类体系覆盖了国民经济的 20 个门类、97 个大类，上市公司根据其主营业务被归入相应的行业类别中。对于上市公司数量众多、业务复杂的制造业（门类 C），体系还特别设置了次类作为辅助，以便更精细地反映行业内部结构。

该分类体系的依据是客观的量化规则与专家判断相结合。具体而言，体系设立了三项递进的量化规则：首先，当公司某一类业务的营业收入比重大于或等于 50%时，原则上直接划入该业务所属行业。其次，在满足前一条件的同时，若另一类业务的营收占比超过 30%且其毛利润占比超过 50%，公司可申请划入后者。最后，当公司无任何一类业务的营收占比达到 50%时，则需综合比较各业务的营收与毛利润占比进行判定，若仍无法确定，则提交专家委员会根据公司实际经营状况进行最终裁定，或归入综合类。

（二）申万行业分类

申万行业分类由申万宏源研究所制定并持续维护。自 2003 年正式发布第一版以来，申万行业分类迄今已完成了五次修订，最新的两次修订分别发生在 2014 和 2021 年。2021 版申万行业分类有 31 个一级行业，134 个二级行业和 346 个三级行业。

在分类依据上，申万行业分类始终坚持“盈利驱动、估值聚类、物理形态和使用习惯”四大核心原则。其判定逻辑与基于经济统计的“管理型分类”存在本质区别，更侧重于上市公司的经营实质与市场表现。具体而言，分类过程中主要考察上市公司各项业务的营业收入与毛利润来源结构，优先将公司划入对其利润贡献最大、业务关联度最高的行业类别中。

（三）万得（Wind）行业分类

万得行业分类是由万得信息技术股份有限公司自主研发的投资型行业分类标准，自2004

年正式推出，最新版（2024 年 11 月）采用四级层级架构，共划分为 11 个一级行业、36 个二级行业、82 个三级行业和 169 个四级行业。与申万分类侧重盈利贡献不同，万得行业分类以全球行业分类系统（GICS）为基础进行本土化调整，其核心特征在于“以产品为核心、以数据为驱动”的分类逻辑。

（四）其他行业分类标准

中信行业分类是由中信证券研究部编制的投资型分类标准，自 2010 年发布，最新版（2020 年）设 30 个一级行业、109 个二级行业和 285 个三级行业，其核心特征是以上市公司收入与利润来源结构为判定依据。中证行业分类由中证指数有限公司编制，采用四级层级架构，共 11 个一级行业，其核心特征在于与指数编制的深度绑定，通过在宽基指数基础上叠加行业筛选构建宽基行业指数（如 300 医药指数），为指数化投资及 ETF 产品设计提供了直接依据。恒生行业分类由恒生指数公司编制，共 12 个一级行业，主要服务于港股市场，随着内地与香港市场互联互通的深化，该分类在 A 股与港股跨市场比较研究中具有重要应用价值。全球行业分类系统（GICS）由标准普尔（S&P）与摩根士丹利资本国际（MSCI）于 1999 年联合编制，采用四级分类架构，共 11 个行业板块和 163 个子行业，以企业的主营业务为核心，侧重全球投资视角下的行业可比性。

（五）目前分类标准的局限

尽管上述分类标准在投资框架中各具价值，但其共同面临着标准更新严重滞后于产业高速更迭的普适性挑战。从具体逻辑来看，不同标准的局限性侧重各异：中上协行业分类因严格遵循营收比例的原则，往往难以捕捉跨界经营企业的创新增长点或软科技属性；申万、中信等民间投资型分类虽然更贴近市场风格，但在分类底层的打分权重与判定细则上往往缺乏公开透明度，导致研究者难以复现其逻辑；而万得等旨在对接 GICS 等国际标准的体系，在追求全球可比性的过程中，时常会削弱对白酒、新能源产业链等具有中国特色支柱产业的细分刻画，导致在本土化应用中出现“水土不服”的现象。这种静态且标准不一的分类现状，已逐渐难以满足当前对产业深度分析及动态行业趋势研判的精准需求。

三、文献综述

（一）上市公司行业分类

行业分类是金融经济学与实证会计研究的基石性工具。其重要性首先体现在对公司基本面差异的解释力上。McGahan 和 Porter（1997）的研究表明，行业效应能够解释公司利润差

异中相当可观的部分，这一效应在零售、服务、运输等非制造业中尤为显著。Alford（1992）的研究证实，基于行业分类来选择可比公司，能够最显著地降低市盈率估值模型的误差，确立了行业分类在界定相似企业时的核心地位。在公司财务领域，Frank 和 Goyal（2009）在系统检验资本结构的决定因素时发现，“行业杠杆率中位数”是解释公司间杠杆率横截面差异的最核心基准因素之一。此外，在资产定价研究中，Moskowitz 和 Grinblatt（1999）揭示了显著的“行业动量”效应，证明个股层面的动量收益在很大程度上是由行业整体的回报动量所驱动的。在中国市场，郭鹏飞与孙培源（2003）的研究同样证实了行业因素的重要性。他们发现，中国上市公司存在最优资本结构，行业是重要影响因素之一。段丙蕾等（2022）在中国市场也发现了行业关联回报率在月度上显著。这些研究表明，无论在美国还是中国市场，行业分类都是理解公司财务、回报率等特征不可或缺的分析工具。

在美国市场，主流行业分类体系大致可分为三类：一是以生产过程为导向的政府统计型分类，包括标准行业分类系统（SIC）和北美行业分类系统（NAICS），前者自 1937 年创立，依据企业的主要产出和生产工艺，以四位数层级结构对经济活动进行划分，长期作为学术研究和监管实践的核心分类工具；后者于 1997 年由美国、加拿大和墨西哥联合推出以替代 SIC，在沿用生产导向逻辑的基础上采用六位数编码体系，扩展了对信息技术、专业服务等新兴行业的覆盖。二是以资本市场为导向的商业型分类，其代表为全球行业分类标准（GICS），由 MSCI 与标准普尔于 1999 年联合开发，以企业的主营业务收入来源为首要分类依据，构建了"行业板块、行业组、行业、子行业"的四级体系，旨在更精准地反映投资者视角下的公司相似性。三是学术研究中广泛使用的 Fama-French 行业分类，其本质上是对 SIC 代码的再归并，由 Fama 和 French 根据研究需要将四位数 SIC 代码映射为不同粒度的行业组合，便于资产定价和投资组合研究中的行业控制。

然而，上述分类体系，尤其是长期占据主导地位的 SIC，其固有缺陷已被大量文献所揭示。Clarke（1989）较早指出，SIC 代码在界定经济市场边界时存在显著偏差，同一 SIC 代码下的企业可能在产品特征、竞争格局和盈利模式上存在实质性差异。Kahle 与 Walkling（1996）发现，Compustat 和 CRSP 两大主流数据库中同一家公司的 SIC 代码在两位数层面存在超过 36%的不一致，在四位数层面更是高达近 80%的不一致，这种数据源之间的分类差异可能导致基于行业匹配的实证研究产生系统性偏误。Bhojraj、Lee 与 Oler（2003）在对 SIC、NAICS、GICS 和 Fama-French 行业分类四大体系的系统性比较中发现，GICS 在解释股票收益共变性、估值倍数的横截面差异以及增长率预测等方面均显著优于其他三种分类，

而 SIC、NAICS 与 Fama-French 分类之间的表现差异则相对有限。这一发现表明，以生产过程为导向的政府统计型分类在捕捉资本市场中公司相似性方面的能力存在明显不足。此外，Hoberg 和 Phillips（2016）指出，包括 SIC 和 NAICS 在内的固定行业分类体系极少对业务发生变化的公司进行重新分类，并且它们不允许行业本身随时间演变。

相较于美国市场 SIC/NAICS 分类的上述问题，中国市场的行业分类体系面临的挑战更为复杂和突出。首先，中上协（原证监会）、申万、万得、中信等多套标准并行，分类逻辑各异，导致研究结果缺乏可比性和可复制性；其次，以中上协为代表的官方管理型标准服务于国民经济统计，其按生产活动物理属性划定行业边界的方式难以反映资本市场的估值逻辑，且更新机制滞后，无法及时覆盖新能源汽车、人工智能等新兴业态；最后，申万等投资型标准虽然市场适应性更强，但其分类过程依赖人工判断、缺乏透明度，且仍采用按主营业务收入占比归类的静态单行业划分方法，无法有效刻画多元化经营和跨界转型企业的真实业务结构。

面对传统行业分类体系的局限，学者们开始探索基于公司披露文本的新型分类方法。Hoberg 和 Phillips（2010）开创性地提出了基于文本分析的产品市场相似性度量方法，通过分析公司 10-K 年报中的产品描述文本，计算公司两两之间的文本相似度，从而识别产品市场上的竞争关系。Hoberg 和 Phillips（2016）正式提出了“基于文本的网络行业分类”（Text-based Network Industry Classification，简称 TNIC）。TNIC 的核心创新在于：它不是将公司归入离散的行业类别，而是构建一个包含所有公司两两相似度的全连接网络。研究者可以根据研究需要设定相似度阈值，只有当两家公司的文本相似度高于阈值时，才被视为同一行业。这一方法的优势在于：其一，它每年基于最新的年报文本重新计算，实现了行业分类的动态更新；其二，它不再依赖预设的行业定义，而是让数据自身揭示公司之间的相似结构；其三，它保留了相似度的连续信息，为研究者提供了更丰富的分析维度。

国内学术界对行业分类方法的系统性研究起步较晚，且现有成果主要集中在实务层面的标准梳理和简单的分类比较上，缺乏对分类方法本身的深入理论探讨和创新。与美国学术界围绕 SIC、NAICS、GICS 和 TNIC 等分类体系展开的大量实证比较研究不同，中国学者更多地将行业分类视为一个既定的外生工具，直接将其应用于公司金融、资产定价等研究场景中，而较少审视分类体系本身的质量和有效性。相比于 Hoberg 和 Phillips（2016）基于美国市场构造的动态、数据驱动的 TNIC 分类方法，中国市场目前尚缺乏一套基于数据驱动、方法公开、能够逐年动态更新的行业分类体系。这种方法论上的空白使得中国市场的行业分类

研究在很大程度上仍停留在对现有标准的被动使用阶段，缺乏对分类质量本身的系统性审视和改进。

本研究从两个方面贡献于上述工作。首先，本研究构建了一套数据驱动的，准确、动态、可拓展的三级行业分类数据集，填补了目前中文学术界暂无上市公司行业分类体系的空白。其次，本研究通过多种度量检验，系统地将本文提出的分类体系与申万、万得和中上协等主流标准进行了定量比较，为中国市场行业分类质量的评估提供了可复制的方法框架。

（二）金融文本分析与大语言模型

金融文本分析是利用自然语言处理（NLP）技术从非结构化的金融文本数据中提取有价值信息的研究方法。随着金融市场中新闻报道、公司年报、分析师报告、社交媒体等文本数据的爆发式增长，文本分析已成为金融经济学研究的重要方法论工具（沈艳，2019；洪永淼、汪寿阳，2021）。早期的金融文本分析主要依赖基于词典的方法。Tetlock（2007）利用《华尔街日报》专栏文本中的悲观情绪词频构建了媒体情绪指标，发现该指标能够显著预测股市的短期下行趋势和交易量变化。Loughran 与 McDonald（2011），姜富伟等（2021）针对金融文本的特殊性，分别构建了专门适用于英文和中文金融文本的情感词典。

然而，基于词典的方法忽略了词语的上下文信息和语义关系。随着深度学习技术的发展，词嵌入（word embedding）方法实现了对文本语义的分布式表示。Mikolov 等（2013）提出的 Word2Vec 模型能够将词语映射到低维连续向量空间，使得语义相似的词语在向量空间中彼此接近。Le 与 Mikolov（2014）进一步提出了 doc2vec 模型，将词嵌入的思想从词级扩展到文档级，能够为变长文本生成定长的向量表示。这两项技术为金融文本分析带来了新的可能性。在行业研究领域，Hoberg 和 Phillips（2025）引入了基于 doc2vec 嵌入模型的公司业务范围度量方法，胡楠等（2020）采用“种子词+Word2Vec 相似词扩充”的方法，基于年报文本对企业竞争战略进行度量。

以 GPT 系列为代表的生成式大语言模型的快速发展，进一步拓展了金融文本分析的应用边界（洪永淼、汪寿阳，2024）。值得注意的是，大语言模型在金融文本分析中的应用可以大致分为两类技术路径：

一是利用生成式模型的文本理解和推理能力，直接对金融文本进行情感判断、信息提取和分析推理。例如，Siano（2025）利用微调后的 Bert 模型捕捉盈余公告中的复杂信息，显著提升了对短期股票异常收益变化的解释力度 ；陆瑶等（2025）结合 FinBERT 和 GPT、ChatGLM 等大模型，基于上市公司年报文本，构建了多层次的企业数字技术风险暴露指标 ；

Li 等（2026）使用 Bert 与 GPT，对分析师报告、盈余电话会议等进行文本分析，不仅精准识别并分类了企业文化类型，还提取了关于企业文化的因果关系逻辑并构建了知识图谱。

二是利用模型的文本嵌入能力，将文本映射为语义向量，进而通过向量间的相似度度量支持下游任务。文本嵌入方法的核心优势在于：它能够将大语言模型预训练过程中积累的丰富语义知识提炼为一个定长的稠密向量，与 Word2Vec 和 doc2vec 等浅层模型相比，基于预训练语言模型的文本嵌入能够学习到远比浅层神经网络更丰富的语法和语义知识，从而生成信息量更密集的稠密向量（Devlin 等，2019）。例如，Breitung 和 Müller（2025）运用 GPT-3 生成历史业务描述，并利用文本嵌入模型处理了全球 63000 多家上市公司的文本数据，首次构建了动态的全球商业网络。

在上述研究基础上，本文将大语言模型的文本嵌入能力应用于中国 A 股上市公司年报 MD&A 文本，通过语义向量化与层次聚类相结合的方法构造行业分类数据集，为文本数据和基于大模型的嵌入技术在金融领域的应用开辟了新的视角和探索路径。

四、数据来源及分类体系构建

（一）样本选择与数据来源

本文使用中国 A 股上市公司年度报告中的“管理层讨论与分析”（MD&A）章节文本构造行业分类体系。样本时间跨度为 2007 至 2023 年，覆盖沪深两市全部 A 股上市公司的年度报告，共计 52702 份。

本研究使用企业年报中的 MD&A 章节文本构造行业分类体系基于以下两个原因：第一，根据《公开发行证券的公司信息披露内容与格式准则第 2 号——年度报告的内容与格式》中二十一至二十六条的相关规定，上市公司需要在年报的 MD&A 章节中详细披露公司从事的业务情况、行业情况、核心竞争力以及报告期内主要经营情况等内容，这一部分业务信息正是我们行业分类的核心依据。相较于年报全文，MD&A 部分的业务信息更集中，并且噪音更少。第二，基于 MD&A 文本匹配相似公司的做法在国内外得到诸多应用，比如，徐巍等（2025）使用 MD&A 部分“报告期内从事的主要业务和产品”相关内容的文本相似度来确定上市公司竞争对手。

本研究选取 2007 年作为样本起始年份，主要源于 2007 年 1 月 1 日正式施行的新《企业会计准则》及配套信息披露规则。新准则体系强制要求上市公司年报中标准化披露主营业务、行业及经营范围，显著提升了文本信息的规范性与丰富度；同时，新制度强调会计信息可比

性并清晰界定主营业务边界，有效消除了因会计制度差异导致的语义噪音。鉴于 2007 年前年报披露标准不一，选取 2007 年为起点能够确保研究期间内信息披露制度环境的一致性，从而保证基于年报文本构建的行业分类体系具有可靠的数据基础与研究结论的稳健性。

本研究其余行业分类数据来自于万得（Wind）数据库，分析所使用的市场和公司财务数据来自国泰安（CSMAR）数据库。

（二）分类体系构建流程

本文参考 Hoberg 和 Phillips（2016，2025）的文本相似度行业分类框架，采用文本嵌入大模型对各公司 MDA 章节文本进行语义向量化，并通过层次聚合聚类方法构建三级行业分类体系。在得到分类结果后，本文借助大语言模型对分类结果进行行业命名。通过两阶段提示词策略，为每个行业类别赋予符合中国 A 股市场惯例的行业名称，从而构建出一套同时具备文本数据驱动特性与经济语义可解释性的多层次行业分类体系。以下是具体的步骤：

1. 文本嵌入与归一化

本研究首先对 52702 份 MD&A 进行文本嵌入向量化处理。文本嵌入向量化是将非结构化的文本映射为高维、稠密的实数向量，向量方向表示了文本的语义。本研究首先将所有 MD&A 文本去除全部空白字符，并且切分为最大长度$L_{max} = 6500$个字符的文本块。[1]在切分文本块时，分块算法在候选窗口内寻找最后一个中文句号作为切分点，以保留语义完整性；若候选窗口内不存在句号，则在第 6500 个字符处进行切分。设第i篇公司年份文档d_i经分块后得到K_i个文本块，记为 $\{c_{i,1}, c_{i,2}, \dots, c_{i,K_i}\}$，其中每个块满足 $|c_{i,k}| \leq L_{max}$。

获取文本块后，本研究使用 Qwen-text-embedding-v4 文本嵌入模型对各文本块进行语义向量化。该模型输出向量为 2048 维。为增强嵌入对行业语义的捕捉能力，本文在调用模型时加入任务指令（Prompt）：

为所提供的公司年报“管理层讨论与分析（MD&A）”部分的文本生成一个语义嵌入。该嵌入应当概括公司的核心商业模式，运营重点及其经济活动。

设文本块$c_{i,k}$对应的嵌入向量为$v_{i,k} \in \mathbb{R}^{2048}$，则该嵌入向量由模型$f_{embedding}$生成：

$$v_{i,k} = f_{embedding}(c_{i,k}; Prompt) \tag{1}$$

对于同一企业年份(s,t)（其中s为股票代码，t为年份），若其 MDA 文本被分割为$K_{s,t}$个

[1] 所使用的嵌入大模型 Qwen-text-embedding-v4 最大输入长度为 8000tokens，在实际研究中，切分为 6500 字符的最大文本块不会出现文本 tokens 数量超过上下文限制的情况。

文本块，则通过对各块嵌入向量取算术平均值的方式，得到该企业年度的代表性语义向量：

$$e_{s,t} = \frac{1}{K_{s,t}} \sum_{k=1}^{K_{s,t}} v_{s,t,k} \tag{2}$$

其中 $e_{s,t} \in \mathbb{R}^{2048}$为企业$s$在年份$t$的行业语义嵌入向量。整个嵌入流程通过阿里云批处理API接口实现，以满足大规模文本处理的效率要求。

在进行聚类分析之前，对所有企业年度嵌入向量$\mathrm{e}_{\mathrm{s,t}}$实施$L_2$归一化处理，以消除向量模长差异对距离度量的影响：

$$\tilde{e}_{s,t} = \frac{e_{s,t}}{\| e_{s,t} \|_2} \tag{3}$$

归一化后，全部企业年度观测值构成矩阵 $X \in \mathbb{R}^{N \times 2048}$，其中$N$为样本总量 52702。

2. 层次聚合聚类

本研究采用层次聚合聚类方法进行不同 MD&A 的聚类。层次聚合聚类（Agglomerative Hierarchical Clustering）是一种自下而上的贪心策略，通过将每个样本初始化为独立簇，并在迭代中基于特定的距离度量与连接准则，不断合并相似度最高的簇对，最终构建出一个嵌套的树状结构，直至达到预设的簇数量或收敛条件。这一方法完全由数据自下而上驱动，避免了预设行业定义和范围可能带来的先验偏差。本文使用平均链接（Average Linkage）准则衡量簇间距离。具体地，两个簇$\mathcal{A}$与$\mathcal{B}$之间的距离定义为所有跨簇样本对之间欧氏距离的均值：

$$D(\mathcal{A}, \mathcal{B}) = \frac{1}{| \mathcal{A} || \mathcal{B} |} \sum_{x_i \in \mathcal{A}} \sum_{x_j \in \mathcal{B}} \|x_i - x_j\|_2 \tag{4}$$

平均链接方法能够产生较为平衡的聚类结构，适合大规模文本嵌入的聚类场景。在文本分析中，通常嵌入向量距离计算会使用余弦距离。在本文中，我们使用归一化后的欧式距离，是考虑到，在归一化前提下，欧式距离和余弦距离存在以下单调映射关系：

$$\begin{aligned} \| \tilde{e}_i - \tilde{e}_j \|_2 &= \sqrt{\tilde{e}_i^T \tilde{e}_i - 2\tilde{e}_i^T \tilde{e}_j + \tilde{e}_j^T \tilde{e}_j} \\ &= \sqrt{1 - 2cos(\tilde{e}_i, \tilde{e}_j) + 1} \\ &= \sqrt{2\left(1 - cos(\tilde{e}_i, \tilde{e}_j)\right)} \end{aligned} \tag{5}$$

在单点聚合的情况下，归一化后的欧式距离和余弦距离是等价的。在计算两个较大簇的平均距离时，MD&A 文本中可能存在语义相对偏离的“离群文本块”，这些块会产生极大的单对

余弦距离。直接平均余弦距离容易受到这些极端值的影响；而归一化欧氏距离的平方根特性在聚合过程中有效压缩了极端样本对的惩罚权重。这使得聚类算法对 MD&A 文本中的局部噪声更加鲁棒，从而构建出更紧凑、语义凝聚度更高的行业簇。

具体而言，本研究构建了一个自底向上的三级级联聚类架构，并创新性地引入了动态小簇合并机制，以解决传统层次聚合聚类容易产生大量极小簇的缺陷。该聚类流程可拆解为以下三个核心步骤：

（1）第三级聚类。本研究以全体归一化样本向量为输入，直接进行基于欧氏距离的层次聚合聚类，采用平均链接准则，将所有样本聚合为 300 个基础簇。聚类完成后，对所有基础簇执行小簇合并：统计每个簇的样本量，将样本量严格小于阈值$\delta_3 = 5$的微小簇识别为“孤岛簇”。对于每个孤岛簇，计算其质心向量与所有满足$n_k \geq \delta_3$的大簇质心向量之间的欧氏距离，并将该孤岛簇中的所有样本重新分配至最近的大簇，从而保证第三级分区中每个有效簇均具有一定的样本规模，并且这种方式也可以保证不同层级分类的完全嵌套关系。重新分配后，本研究获得了 271 个第三级簇。

（2）第二级聚类。在构建第二和第一层级分类体系时，本研究不再直接依赖原始样本向量，而是基于上一级的聚类结果构建簇间距离矩阵。具体做法为：在第二（一）级聚类中，提取第三（二）级聚类中任意两个簇，计算这两个簇内所有跨簇样本向量对之间欧氏距离的均值，将其作为这两个簇的距离。随后，以第三（二）级的簇间距离矩阵为输入，继续执行平均链接的聚合，从而严格保证了层次结构的嵌套性。这一方法相当于给第三（二）级每个类，无论大小，都赋予了相同权重，可以有效遏制聚类过程中大类吞并小类的情况。

按照上述方法，在第二级聚类中，基于第三级聚类的结果，构建簇间距离矩阵$D^{(3)} \in \mathbb{R}^{|\mathcal{C}_3|\times|\mathcal{C}_3|}$。矩阵中第$(i,j)$个元素定义为第三级簇$\mathcal{A}_i$与$\mathcal{A}_j$之间所有跨簇样本对欧氏距离的均值，即：

$$D(\mathcal{A}_i,\mathcal{A}_j) = \frac{1}{|\mathcal{A}_i||\mathcal{A}_j|}\sum_{x_p\in\mathcal{A}_i}\sum_{x_q\in\mathcal{A}_j}\|x_p - x_q\|_2 \tag{6}$$

采用平均链接准则，将 271 个第三级簇进一步聚合为 150 个第二级簇。之后同样执行小簇合并：将样本量$n_k < \delta_2 = 30$的簇并入最近的大簇，以确保中间层分区的类别平衡性。

（3）第一级聚类。类比第二级聚类的步骤，基于第二级聚类的结果，重新构建第二级簇间距离矩阵$D^{(2)} \in \mathbb{R}^{|\mathcal{C}_2|\times|\mathcal{C}_2|}$，并以预计算方式输入层次聚合聚类，采用平均链接准则，将所有第二级簇进一步聚合为 50 个一级行业簇。完成后执行小簇合并，阈值$\delta_1 = 300$，以保

证一级类别具有足够的样本规模和统计代表性。

经过上述聚类过程，本研究最终获得了三级分类数量分别为 26、102、271 的三层中国 A 股上市公司分类体系。整个分类体系完全嵌套，即若任意两家上市公司归属于同一三级行业，则它们必然也归属于同一个二级行业和一级行业。

3. 行业语义命名

按照行业分类的惯例和使用上的便利，我们需要对第二步中分出来的行业进行命名。在参考了申万和万得分类体系的命名方式后，考虑到三级行业分类过于详细，仅凭行业名称可能无法显著区分，本研究只对第一和第二级的行业进行了名称命名，对于三级的行业则采用其所属的二级行业名称加上罗马数字后缀以示区分。

本研究创新性采用了基于 LLM 的命名方式，这一方法可以有效避免人工命名带来的偏好偏差，并且可以最大限度捕获同一层级不同行业的区别。由于直接将大量原始文本输入模型进行全局比对命名会导致严重的上下文超限问题，本研究利用大语言模型设计了一套“局部摘要-全局命名”的两阶段提示词策略。[2]该策略依次对第一层级和第二层级分类结果进行处理，具体步骤如下：

（1）一级行业摘要与命名。首先，本研究采用具备长上下文处理能力的 Qwen-Long 模型，对各个一级行业进行底层文本特征提炼。该模型的最大输入长度为 1000 万 tokens，足够覆盖本研究的需要。为平衡样本代表性与模型 API 容量限制，针对每个行业，随机抽取不超过 1500 篇企业年度 MD&A 文档，并截取每篇文档的前 1000 个字符以捕捉核心业务陈述。在提示词中，设定模型为“专业行业业务分析师”，要求其基于输入的簇内截断文本，深度提炼并输出一份详尽的行业业务画像总结（涵盖核心产品、关键技术及服务对象等）。

在获取所有一级行业的业务摘要后，本研究转用 Qwen3-Max 模型进行全局统筹命名。为了保证各行业名称的互斥性，本研究将所有行业的业务摘要整合为单一输入，要求模型进行全局对比分析。通过在一轮对话中输入所有数据，有效避免了孤立命名导致的语义重叠。约束条件严格限制输出名称必须符合中国 A 股市场通用术语（如“基础化工”、“食品饮料”），长度限制在 2 至 6 个中文字符，并强制以标准 JSON 格式输出映射字典。

（2）二级行业摘要与命名。二级行业的命名逻辑与第一层级相似，同样也是先提炼各个二级行业的特征，再进行统一命名。但为了确保分类体系在经济语义上的严格嵌套与从属

[2] 完整的提示词详见附录部分。

关系，本研究的命名并非对所有二级行业进行一次性混合处理，而是采取了以一级行业为基本单元的“分批命名”策略，并在提示词工程中显式引入了一级行业信息。

所有二级行业在生成业务摘要时，都会先被其所属的一级行业分组。除了向 Qwen-Long 模型输入该一级行业下所有二级行业抽样获得的 MD&A 文本片段外，本研究还将该一级行业的名称及对应的行业业务总结作为先验背景一并置于提示词中。这使得模型在提炼二级行业特征时，能够紧密结合其在一级行业宏观图景下的具体定位，产出具备层级关联性的业务画像。

在随后的二级行业命名环节，本研究同样以一级行业为基本单元进行分组批处理。对于每个一级行业，将其背景信息与其下属所有二级行业的业务摘要同时输入到 Qwen-Max 模型。提示词明确要求模型结合一级行业背景，生成体现强从属或细分关系的名称（例如：在一级行业为“汽车”的背景下，即使某二级行业涉及机械加工，也应侧重命名为“汽车零部件”）。名称长度被放宽控制在 4 至 10 个中文字符，以适应细分行业更具体的表述需求。最后，算法通过全局遍历进行名称重复性检验，确保整套分类体系中 102 个第二层级行业名称的绝对唯一性与精准映射。

五、实证检验及分析

（一）LLM 分类概述

经过上述构建流程，本研究最终形成了一套三级嵌套的中国 A 股上市公司行业分类体系（以下简称 LLM1-3 级分类体系），具体包括 26 个一级行业、102 个二级行业与 271 个三级行业。从整体颗粒度来看，该体系在一级层面的类别数量（26 个）与主流分类标准（申万 31 个、万得一级 11 个、万得二级 36 个）大体相当；在二级层面（102 个）对标申万二级（134 个）与万得三级（82 个），颗粒度处于两者之间；而三级层面（271 个）与申万三级（346 个）规模相近，相较于万得四级（169 个）层次更为精细。

图 1 展示了一级分类体系下 26 个行业每年平均公司数量分布。在一级行业中，超过 70% 的行业都有不少于 50 个公司，其中最小的两个行业年平均有 18、19 个公司。

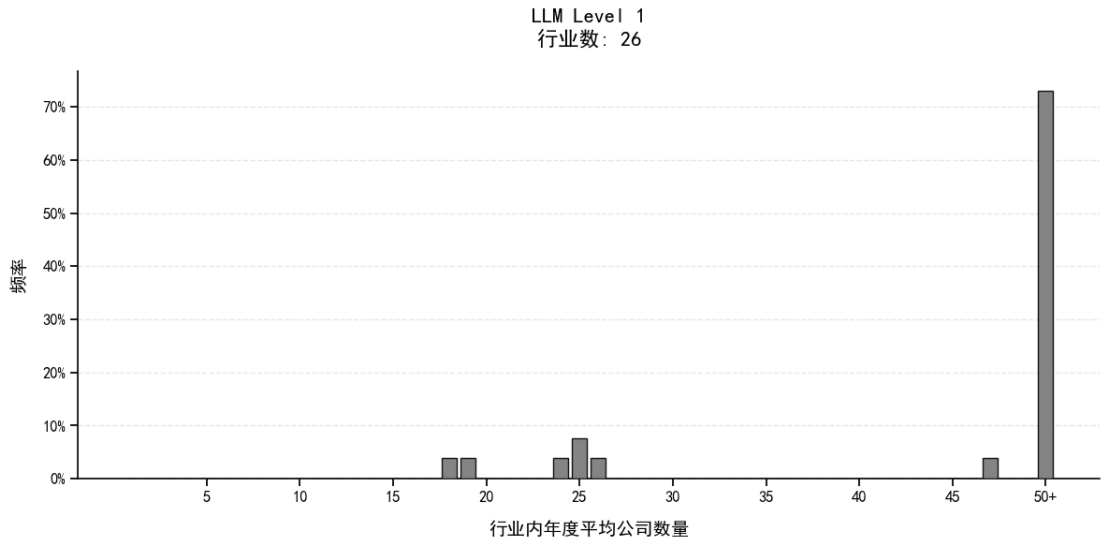

图 1　LLM 一级分类下行业平均数量分布

注：图 1 展示了本文构建的 LLM 一级行业的平均公司数量分布情况。横轴表示行业内年度平均公司数量，纵轴表示处于各区间内的行业数量占全部一级行业的百分比。

表 1 展示了 26 个一级行业的名称、平均公司数量以及 2007、2015 和 2023 年的各行业公司数量情况。这 26 个一级行业涵盖了中国 A 股市场的主要经济活动领域，包括：高端装备、食品饮料、医药生物、电子元件、软件服务、农林牧渔、基础化工、种子农业、交通运输、电力设备、公用事业、纺织服装、文化传媒、建筑材料、商业零售、房地产、综合转型、交运能源、金融服务、家电部件、旅游酒店、轨道交通、造纸包装、高速公路、石油化工和环保水务。

从平均规模来看，制造类行业（如高端装备、基础化工、电子元件等）平均容纳公司数量超过 300 家，而部分细分领域（如轨道交通、种子农业等）平均不足 25 家，行业间规模差异显著。时间上来看，2007 年至 2023 年间，中国 A 股上市公司数量整体呈现大幅增长，但各行业表现分化明显。绝大多数行业公司数量显著增加，其中电子元件、软件服务、文化传媒、环保水务和电力设备等行业扩张最为迅速。然而，部分行业增长缓慢甚至出现萎缩：综合转型行业公司数量从 2007 年的 294 家骤降至 2023 年的 3 家，可能由于企业逐步完成转型或行业整合；房地产行业在 2015 年达到 147 家的高点后回落至 122 家，呈现先升后降的态势；商业零售和旅游酒店行业公司数量基本持平。此外，一些细分领域如种子农业、交运能源、高速公路等，公司数量长期维持在较低水平，2023 年均不足 30 家，增长幅度有限，显示出这些行业的高度集中或市场容量限制。总体而言，表 1 的数据揭示了中国 A 股市场

行业结构的动态演变，高新技术产业和先进制造业快速扩张，而部分传统行业则面临调整或增长瓶颈。

表 1　LLM 一级分类下 26 个行业名称与公司数量

行业编号	行业名称	平均公司数量	2007 年公司数量	2015 年公司数量	2023 年公司数量
1	高端装备	363.7	111	320	715
2	食品饮料	63.5	24	58	110
3	医药生物	240.5	96	211	449
4	电子元件	324.4	71	255	766
5	软件服务	269.8	60	240	504
6	农林牧渔	84.8	31	87	134
7	基础化工	309.5	162	313	464
8	种子农业	24.9	20	25	27
9	交通运输	78	48	75	106
10	电力设备	138.4	34	133	242
11	公用事业	126.6	88	126	171
12	纺织服装	176.1	97	158	265
13	文化传媒	99.9	16	105	158
14	建筑材料	130.1	61	131	196
15	商业零售	55.4	45	59	52
16	房地产	139.6	116	147	122
17	综合转型	95.1	294	18	3
18	交运能源	19.1	17	20	21
19	金融服务	66.2	29	52	110
20	家电部件	46.6	18	45	85
21	旅游酒店	24.1	19	27	25
22	轨道交通	24.8	9	23	48
23	造纸包装	73.1	36	70	105
24	高速公路	17.9	15	18	19
25	石油化工	25.8	13	27	34

26	环保水务	82.2	17	78	162

注：表 1 展示了本文构建的 LLM 一级分类体系下 26 个行业的名称及公司数量分布情况。其中，行业名称由“行业语义命名”小节所述方法生成，“平均公司数量”为 2007—2023 年各行业公司数量的年度平均值；“2007 年公司数量”“2015 年公司数量”和“2023 年公司数量”分别列示了样本区间起点，中间节点和终点年份各一级行业的公司数量。

图 2 和图 3 进一步展示了二级和三级分类体系下各行业的平均公司数量分布。从二级分类来看，行业平均公司数量分布呈现较为集中的特征：超过 20%的二级行业年平均公司数量超过 50 家，其余多数行业分布在 5 至 50 家之间，仅有极少数细分类别的公司数量相对偏低，没有单公司行业。三级分类方面，由于行业粒度最细（共 271 个），平均每个三级行业涵盖的公司数量更少，超过 50 家的仅占 5%左右，其中有接近 30%为单公司行业，这一发现也与 Hoberg 和 Phillips（2016）的发现类似，这一聚类方法倾向把独特的公司单独列为一个行业。这也表明，三级分类体系适用于对行业内部差异进行精细刻画的研究场景。

总体来看，三级嵌套分类体系在类别数量与每类样本量之间取得了良好的平衡：一级行业提供了较为宽泛的行业分类，同时能够较好区分各经济活动领域并揭示行业的演变过程，二级分类提供了适中的行业划分，三级分类则提供了更细颗粒度的行业结构。

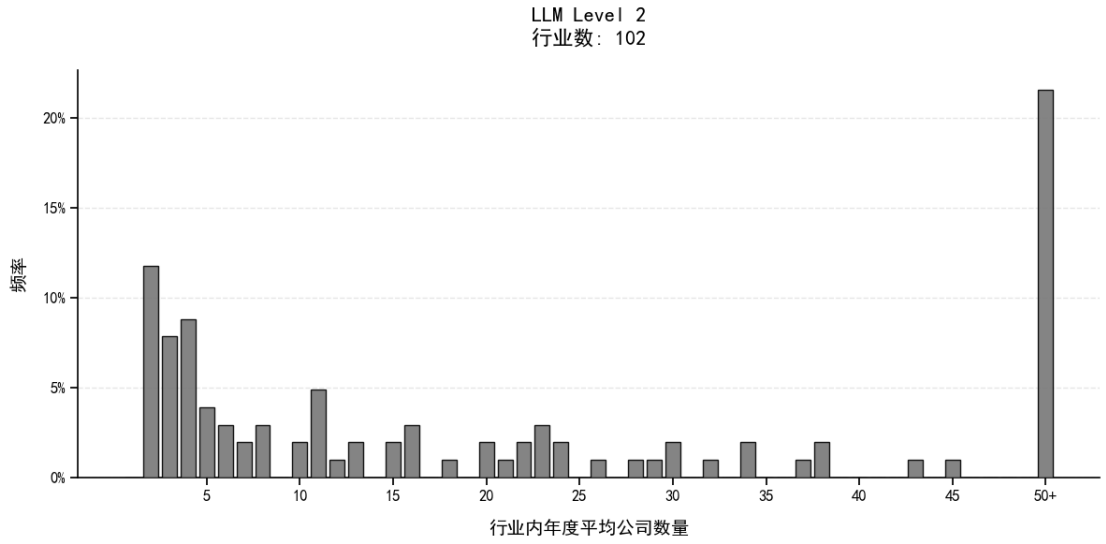

图 2　LLM 二级分类下行业平均数量分布

注：图 2 展示了本文构建的 LLM 二级行业的平均公司数量分布情况。横轴表示行业内年度平均公司数量，纵轴表示处于各区间内的行业数量占全部二级行业的百分比。

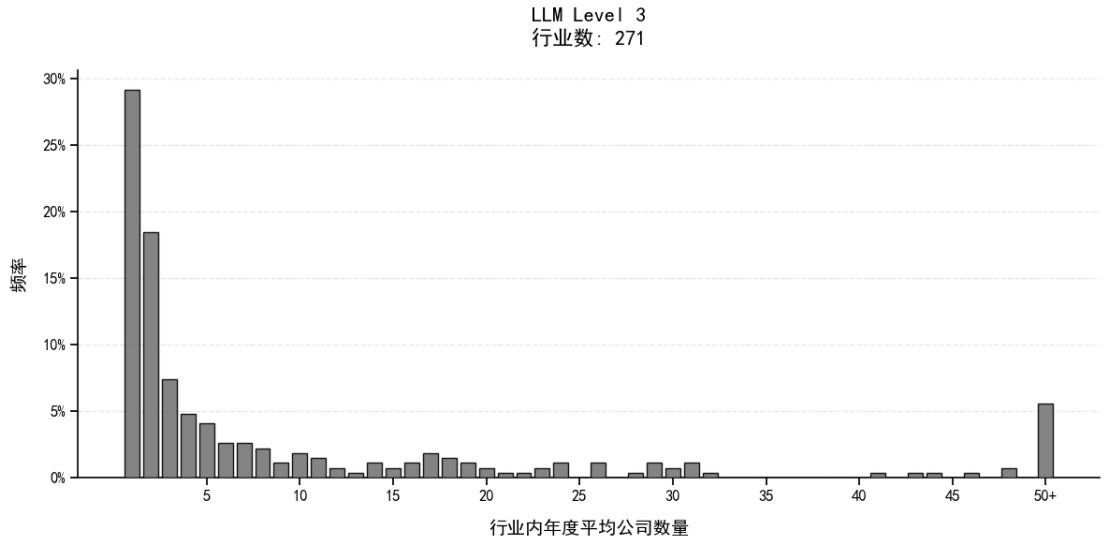

图 3　LLM 三级分类下行业平均数量分布

注：图 3 展示了本文构建的 LLM 三级行业的平均公司数量分布情况。横轴表示行业内年度平均公司数量，纵轴表示处于各区间内的行业数量占全部三级行业的百分比。

（二）行业间差异性分析

衡量分类质量的一个关键维度是行业间的差异性，一个理想的分类体系应当使不同行业在财务特征上呈现出尽可能大的差异。参照 Hoberg 与 Phillips（2016）的做法，本研究计算了各分类体系下不同行业在四个核心财务指标上的标准差，以此衡量分类体系的行业间区分能力。四个财务指标分别为：营业利润率（OpMargin）、资产回报率（ROA）、营业收入增长率（RevGrowth）和资本支出增长率（CapxGrowth）。标准差越大，表明各行业在该指标上的分布越分散，即行业间的财务特征差异越显著，分类质量越高。本研究的行业分类每年更新一次，有效期为当年 7 月 1 日至次年 6 月 30 日。LLM 分类基于年报文本构建（年报于每年 4 月 30 日前披露），因此我们于每年 7 月 1 日确定申万、万得及中上协的行业分类，并将该时点的分类结果与当年披露的对应年报的 LLM 分类共同作为该公司未来一年的行业归属。为避免前瞻性偏差，所有财务指标均选取每年年底的数据，确保其至少滞后于分类结果半年以上。[3]

[3] 为进一步确保财务数据的准确性，在本节及后续分析中，本研究剔除了t年拥有行业分类但在$t+1$年 4 月 30 日之后才首次上市的公司-年份观测。此类公司在财务指标观测日（t年 12 月 31 日）尚未上市，其t年度财务数据仅在 IPO 时通过招股说明书首次公开披露，而非在$t+1$年 4 月 30 日（年报法定披露截止日）前通过上市公司年报渠道公开可得。以$t+1$年 4 月 30 日（年报法定披露截止日）为判断节点，可确保样本中所有公司的t年度财务数据均

表2汇报了2007至2023年全样本期间，基于本研究构建的三级LLM分类体系（LLM1—LLM3）与申万三级分类（SW1—SW3）、万得四级分类（WIND1—WIND4）及中上协分类（CAPCO）在上述指标上的等权与行业内公司数量加权标准差。[4]

表2　各分类体系行业间财务指标标准差

分类	类数	OP 等权	OP 加权	ROA 等权	ROA 加权	Rev 等权	Rev 加权	Capx 等权	Capx 加权
LLM1	26	0.113	0.081	0.025	0.022	0.156	0.123	0.706	0.485
SW1	31	0.064	0.056	0.022	0.02	0.135	0.119	0.566	0.452
WIND1	11	0.097	0.056	0.019	0.016	0.126	0.105	0.453	0.344
WIND2	35	0.079	0.061	0.022	0.019	0.149	0.125	0.645	0.467
LLM2	98	0.188	0.104	0.044	0.029	0.284	0.159	1.626	0.764
SW2	125	0.102	0.081	0.037	0.029	0.198	0.157	0.951	0.676
WIND3	69	0.098	0.075	0.031	0.023	0.198	0.141	1.1	0.579
LLM3	221	0.266	0.132	0.066	0.035	0.381	0.194	2.713	1.112
SW3	290	0.131	0.1	0.048	0.035	0.255	0.188	1.511	0.964
WIND4	131	0.132	0.083	0.042	0.027	0.236	0.158	1.468	0.726
CAPCO	123	0.146	0.083	0.051	0.026	0.309	0.149	2.081	0.737

注：本表汇报了不同分类标准下不同财务指标的标准差。LLM1—LLM3 分别指本研究构建的一级、二级和三级分类；SW1—SW3 为申万一至三级分类；WIND1—WIND4 为万得一至四级分类；CAPCO 为中上协分类。"类数"为各分类标准历年行业数量均值。"等权"指行业内先对指标算均值后，在年份-行业层面计算标准差；"加权"指基于行业内公司数量加权后计算标准差。所有指标在全样本 1%和 99%分位数处缩尾处理。

为了便于比较，本节以及之后的分析都把不同的分类标准按照行业数量大致划分为三个层级：第一层级有 LLM1，SW1，WIND1 和 WIND2；第二层级有 LLM2，SW2 和 WIND3；第三层级有 LLM3，SW3，WIND4 和 CAPCO。其中 WIND4 和 CAPCO 虽然数量上相较于 LLM3 和 SW3 偏少，但其都代表了该分类标准下颗粒度最细的分类，因此放在第三层级统

在该日期前通过标准信息披露渠道向市场公开，从而保证所用数据在相关时点均为真实可得的公开信息。

[4] 由于中上协分类中，不同行业分类颗粒度不同。因此此处的分类是选取了每个行业最细的分类进行统计（例如，金融业只有门类和大类二级分类，而制造业有门类-大类-中类三级分类，此处就是把金融业的每个大类视为一类，制造业的每个中类视为一类）。

一比较。

从表 2 可以观察到几个显著规律。第一，在同等类别数量粒度下，本研究构建的 LLM 分类体系在多数指标上均优于同级别的申万、万得和中上协分类。以营业利润率（OP）为例，在等权算法下，LLM 分类的标准差普遍高于其他体系：LLM1（0.113）高于 SW1（0.064）、WIND1（0.097）和 WIND2（0.079）；LLM2（0.188）显著超过 SW2（0.102）与 WIND3（0.098）；LLM3（0.266）亦明显高于 SW3（0.131）、WIND4（0.132）及 CAPCO（0.146）。第二，各分类体系的行业间差异性随分类颗粒度的细化而提升，这是符合预期的规律性结论，但 LLM 体系在每个层级上的提升幅度均更为显著。以资本支出增长率等权标准差为例，从 LLM1（0.706）到 LLM2（1.626）再到 LLM3（2.713），呈现出近乎翻倍式的递增，而 SW 体系从 SW1（0.566）到 SW3（1.511）的增幅相对温和。第三，数量加权标准差普遍低于等权标准差，这表明规模较大的行业内部财务指标相对集中，而规模较小的细分行业财务特征差异更为突出。

为确保上述结论的稳健性，本研究设计了两组稳健性检验。稳健性检验一（表 3）对 2007 至 2023 年逐年进行 1%和 99%分位数缩尾处理并计算各年标准差，最终汇报各年标准差的时序均值，以避免某一年份的极端情况对整体结果产生过度影响。稳健性检验二（表 4）在逐年缩尾的基础上，进一步将各分类方法的样本取交集后计算标准差，以规避因样本覆盖范围不同而导致的可比性问题。

表 3　财务指标标准差稳健性检验（逐年计算）

分类	类数	OP 等权	OP 加权	ROA 等权	ROA 加权	Rev 等权	Rev 加权	Capx 等权	Capx 加权
LLM1	26	0.112	0.091	0.024	0.022	0.137	0.102	0.742	0.521
SW1	31	0.067	0.059	0.022	0.02	0.116	0.098	0.571	0.48
WIND1	11	0.097	0.056	0.018	0.014	0.111	0.078	0.46	0.331
WIND2	35	0.089	0.065	0.022	0.018	0.137	0.106	0.622	0.487
LLM2	98	0.2	0.126	0.045	0.03	0.3	0.16	1.792	0.901
SW2	125	0.109	0.087	0.037	0.029	0.191	0.146	0.965	0.756
WIND3	69	0.109	0.081	0.031	0.023	0.188	0.127	1.115	0.637
LLM3	221	0.298	0.157	0.068	0.036	0.41	0.206	2.786	1.292
SW3	290	0.14	0.108	0.048	0.036	0.271	0.189	1.717	1.137
WIND4	131	0.154	0.094	0.042	0.027	0.233	0.15	1.558	0.819

CAPCO	123	0.149	0.094	0.046	0.026	0.294	0.149	2.015	0.892

注：同表 2。方法为 2007—2023 年逐年 1%和 99%缩尾后分别计算标准差，再取时序均值。

表 4　财务指标标准差稳健性检验（逐年计算+统一样本）

分类	类数	OP 等权	OP 加权	ROA 等权	ROA 加权	Rev 等权	Rev 加权	Capx 等权	Capx 加权
LLM1	26	0.099	0.077	0.023	0.021	0.128	0.095	0.649	0.455
SW1	31	0.067	0.06	0.022	0.02	0.117	0.098	0.57	0.48
WIND1	11	0.097	0.056	0.018	0.014	0.112	0.079	0.459	0.331
WIND2	35	0.09	0.065	0.022	0.018	0.138	0.107	0.623	0.488
LLM2	96	0.157	0.104	0.042	0.028	0.269	0.144	1.641	0.793
SW2	124	0.109	0.087	0.037	0.029	0.193	0.147	0.961	0.758
WIND3	69	0.109	0.081	0.031	0.023	0.19	0.128	1.117	0.638
LLM3	211	0.218	0.128	0.06	0.033	0.356	0.183	2.448	1.131
SW3	289	0.14	0.108	0.048	0.036	0.274	0.191	1.718	1.138
WIND4	131	0.154	0.094	0.043	0.027	0.238	0.152	1.555	0.819
CAPCO	116	0.15	0.094	0.045	0.026	0.294	0.15	1.834	0.873

注：同表 2。方法为取所有分类方法样本的交集，逐年缩尾处理后计算标准差并取时序均值。

两组稳健性检验的结论与主回归结果高度一致：在控制了年份效应以及样本差异后，本研究构建的 LLM 分类体系在同等颗粒度下仍表现出更高的行业间差异性。综合来看，本研究构建的 LLM 分类体系在行业间差异性维度上具有显著优势，表明该方法能够更有效地将财务特征相异的公司归入不同行业类别。

（三）行业内相似性分析

行业分类质量的另一核心评价维度是行业内的相似性，一个优质的分类体系应当将财务特征相近的公司汇聚于同一行业内。本研究采用行业哑变量回归的 R^2 作为衡量行业内相似性的指标。具体而言，对于每个财务指标（营业利润率 OpMargin、资产回报率 ROA、营业收入增长率 RevGrowth、资本支出增长率 CapxGrowth），本研究在每年将该指标对行业哑变量进行 OLS 回归，并将逐年回归得到的调整后 R^2 计算均值，所得 R^2 均值反映了行业分类能够解释该财务指标横截面差异的比例。R^2 均值越高，表明同一行业内的公司在该指标上越趋同，即分类体系的组内同质性越强。为了避免前瞻性偏差，本节继续使用了滞后半年以上的财务指标。

表 5 报告了各分类体系在四个财务指标上的调整后 R^2 均值。从表 5 可以看出，在同等颗粒度比较下，本研究的 LLM 分类体系在多数指标上具有较高的行业内 R^2。在一级分类层面，除了在收入增长率的平均 R^2 上 LLM1 略低于 WIND2，在其它三个指标上 LLM1 均领先于 SW1，WIND1 和 WIND2 分类。在二级分类层面，LLM2 在营业利润率与资产回报率上仍有更强的解释力，而在收入增长率指标的解释力上略微落后于 SW2，但领先于 WIND3。在三级分类层面，LLM3 同样优势明显，除了资产回报率的平均 R^2 略低于 SW3 外，其他指标均显著优于 SW3，WIND4 和 CAPCO 分类。

表 5　各分类体系行业内财务指标回归 R^2

分类	类数	OpMargin R^2	ROA R^2	RevGrowth R^2	CapxGrowth R^2
LLM1	26	0.064	0.067	0.034	0.009
SW1	31	0.046	0.057	0.034	0.009
WIND1	11	0.032	0.033	0.025	0.006
WIND2	35	0.051	0.046	0.041	0.008
LLM2	97	0.11	0.103	0.054	0.016
SW2	125	0.08	0.103	0.056	0.002
WIND3	69	0.079	0.069	0.051	0.01
LLM3	211	0.144	0.124	0.071	0.031
SW3	289	0.102	0.132	0.061	0.001
WIND4	131	0.095	0.083	0.058	0.011
CAPCO	116	0.079	0.073	0.049	0.018

注：本表汇报了不同分类标准下不同财务指标对行业哑变量的回归 R^2。"类数"为各分类标准历年行业数量均值。表格中汇报的 R^2 为每年各指标对行业固定效应回归的调整后 R^2 的逐年均值，反映行业分类解释财务指标横截面差异的比例。所有财务指标逐年在 1%和 99%分位数处缩尾处理。

表 6 汇报了对上述结果的稳健性检验，处理方法为：在年份-公司的面板数据基础上，控制行业×年份固定效应进行回归，所汇报的是不同财务指标、不同分类标准下该回归的调整后 R^2。

表 6　财务指标回归稳健性检验（交互固定效应）

分类	类数	OpMargin R^2	ROA R^2	RevGrowth R^2	CapxGrowth R^2

LLM1	26	0.059	0.068	0.07	0.014
SW1	31	0.048	0.066	0.069	0.015
WIND1	11	0.034	0.042	0.058	0.012
WIND2	35	0.052	0.057	0.077	0.016
LLM2	97	0.096	0.104	0.095	0.024
SW2	125	0.088	0.113	0.102	0.014
WIND3	69	0.078	0.078	0.09	0.019
LLM3	211	0.125	0.122	0.114	0.046
SW3	289	0.11	0.143	0.113	0.016
WIND4	131	0.09	0.09	0.102	0.022
CAPCO	116	0.08	0.083	0.088	0.028

注：同表 5。表格中汇报的 R^2 为在年份-公司的面板数据基础上，控制行业×年份固定效应回归的调整后 R^2。所有财务指标在全样本 1%和 99%分位数处缩尾处理。

稳健性检验的结论与主回归基本一致，LLM 行业分类体系在三个层级的多数指标上仍维持领先地位。综合行业间差异性与行业内相似性两个维度的证据，本研究构建的 LLM 分类体系在同等颗粒度下能够更好地实现"类内相似、类间差异"的分类目标，具有显著的质量优势。

六、拓展性分析

（一）投资组合构造

行业分类质量的另一重要检验维度来自资产定价领域的“领先-滞后（lead-lag）”效应。Moskowitz 与 Grinblatt（1999）发现，在美国股市中存在显著的行业动量效应，即属于同一行业的公司股票回报之间具有一定的序列相关性，先行公司的回报能够预测滞后公司的未来回报，这一效应本质上源于信息在行业内相似公司之间的传播存在摩擦。在中国 A 股市场，段丙蕾等（2022）同样证实了行业关联回报率在月度上的显著性。若某一分类体系能够更准确地界定“同行业”范围，则基于该分类构造的领先-滞后投资组合应当能够捕获更高的超额回报，从而提供行业分类质量的间接证据。

然而，中国股市的交易机制与美国市场存在显著差异。Du 等（2025）发现，由于 A 股市场“100 股一手”的买入规则，高价股的持有者多为机构投资者，散户比例较低，因而高

价股的个股动量效应更为显著，噪音交易和短期反转对高价股的干扰相对更少。基于这一发现，本研究假设：在高价股子样本中，同行业公司之间的领先-滞后效应同样更为显著，即当焦点公司与关联公司均为相对高价股票时，关联公司的历史回报能够更有效地预测焦点公司的未来回报。

投资组合的具体构造流程如下。首先，在每月月底（t-1 月），从全市场中筛选出收盘价不低于 10 元且流通市值位于市场前 70%的股票作为候选股票池，以排除微市值和低价股的噪音干扰。随后，对于股票池中的每只股票，计算其 t-12 至 t-2 月（即过去 11 个月中排除最近一个月）同行业公司的等权平均累计回报率（排除自身），作为该股票当月的“领先-滞后”特征变量；若某股票在过去 11 个月内任意一个月的同行业有效公司数量少于 5 家，则剔除该观测，以确保同行业参照组的统计可靠性。

在构造多空组合时，参考 Du 等（2025）的做法，本研究采用双重独立排序方法：首先，在 t-1 月底，分别基于收盘价将股票划分为前 10%（高价组）和后 10%（低价组），并基于“领先-滞后”特征将股票独立划分为前 20%（高领先-滞后组）和后 20%（低领先-滞后组）；然后，取上述两组独立排序的交集构建双重分类组合。最终的对冲投资组合为：做多同时属于“高价组”与“高领先-滞后组”的股票，做空同时属于“高价组”与“低领先-滞后组”的股票，考察第 t 月的投资组合收益。上述策略同时考察等权加权和流通市值加权两种组合方式。[5]

[5] 由于中上协（CAPCO）行业分类在 2012 年经历过一次重大调整，所有行业代码都完全更换，调整前后不具有可比性，无法构造投资组合，因此在拓展性分析部分只对比 LLM、SW 和 WIND 三种分类结果。

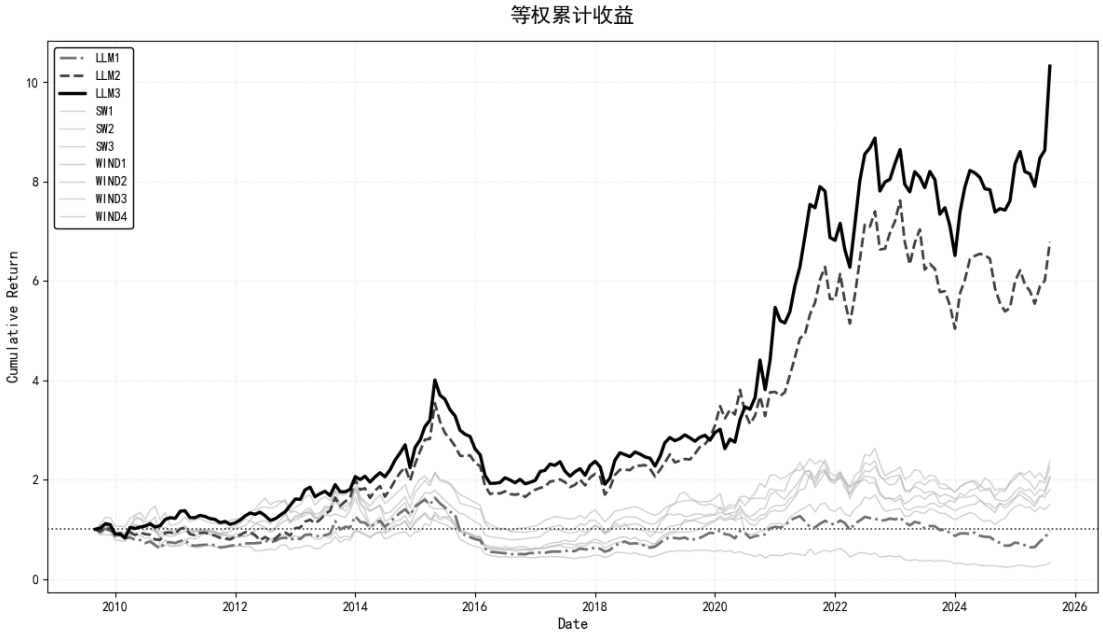

图 4 等权对冲投资组合累计收益

注：该图展示了基于不同行业分类体系构造的领先-滞后对冲投资组合在等权加权方式下的累计收益表现。每月末，在收盘价不低于 10 元且流通市值位于市场前 70%的股票池中，分别按收盘价（前 10% vs 后 10%）和同行业领先-滞后动量（前 20% vs 后 20%）进行双重独立排序，取交集构造多空组合：做多“高价股+高动量”组，做空“高价股+低动量”组。组合每月调整一次，收益率为等权平均。LLM1—LLM3 为本文构建的一至三级分类，SW1—SW3 为申万一至三级分类，WIND1—WIND4 为万得一至四级分类。

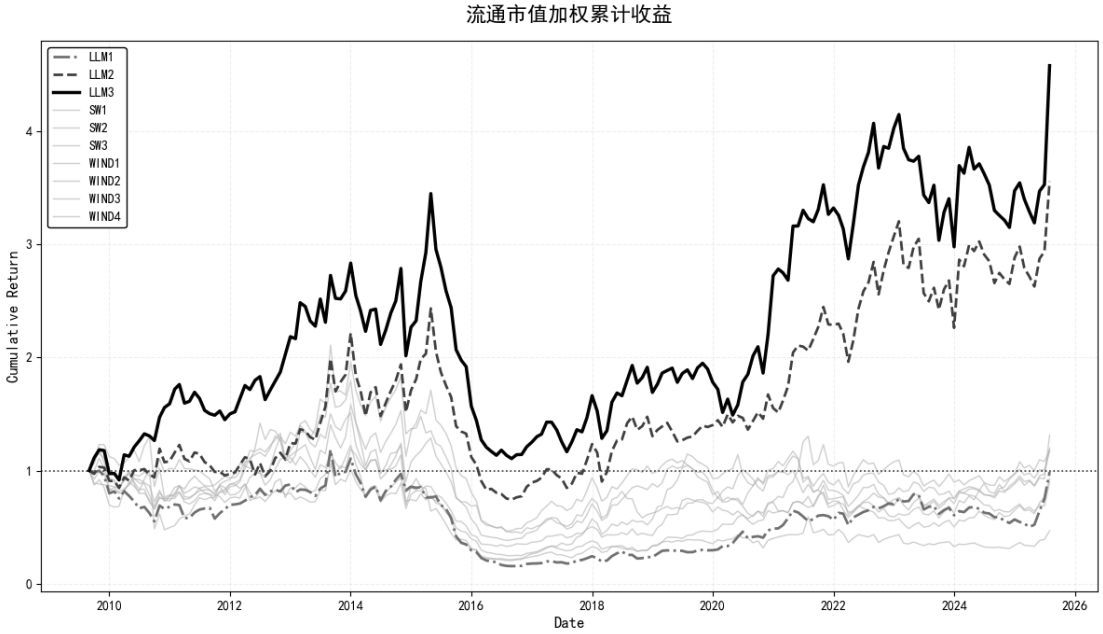

图 5 流通市值加权对冲投资组合累计收益

注：该图展示了基于不同行业分类体系构造的领先-滞后对冲投资组合在流通市值加权方式下的累计收益表现。组合构造方法与图 4 一致，但个股收益按流通市值加权计算组合收益率。其余设定同图 4。

图 4 与图 5 分别展示了等权和流通市值加权对冲投资组合在各分类体系下的累计收益曲线。可以看出，不论是等权还是市值加权下，基于 LLM 二级和三级分类构造的对冲投资组合表现均显著优于其他投资组合。

表 7 与表 8 分别报告了等权和流通市值加权对冲投资组合在各分类体系下的月均回报、T 统计量（经 Newey-West 六阶滞后调整）及 p 值。

表 7　等权对冲投资组合表现

分类方法	平均股票数	平均回报	T-Stat	P-Value
LLM1	40	0.25%	0.42	0.68
LLM2	38	1.29%	2.43	0.02
LLM3	34	1.53%	2.81	0.00
SW1	43	0.64%	1.25	0.21
SW2	37	0.68%	1.34	0.18
SW3	26	0.97%	1.46	0.14
WIND1	44	-0.30%	-0.54	0.59
WIND2	44	0.58%	1.19	0.23
WIND3	42	0.68%	1.41	0.16
WIND4	38	0.49%	0.96	0.33

注：该表展示了等权下对冲投资组合的月度平均回报率与显著性水平。"平均股票数"为每个分类方法下每月平均进入投资组合的多头+空头股票总数；T 统计量经 Newey-West 滞后 6 阶调整。

表 8　流通市值加权对冲投资组合表现

分类方法	平均股票数	平均回报	T-Stat	P-Value
LLM1	40	0.37%	0.48	0.63
LLM2	38	1.03%	1.76	0.08
LLM3	34	1.16%	1.86	0.06
SW1	43	0.27%	0.46	0.65
SW2	37	0.32%	0.54	0.59

SW3	26	0.80%	0.94	0.35
WIND1	44	0.00%	0.01	0.99
WIND2	44	0.46%	0.74	0.46
WIND3	42	0.45%	0.69	0.49
WIND4	38	0.40%	0.65	0.51

注：该表展示了流通市值加权下对冲投资组合的月度平均回报率与显著性水平。"平均股票数"同表 7；T 统计量经 Newey-West 滞后 6 阶调整。

从等权组合结果（表 7）来看，本研究构建的 LLM 二级和三级分类表现尤为突出。LLM2 的等权月均回报为 1.29%（T=2.43，p=0.02），LLM3 为 1.53%（T=2.81，p=0.00），均在统计上显著优异。相比之下，申万体系在二级（SW2：0.68%，T=1.34，p=0.18）和三级（SW3：0.97%，T=1.46，p=0.14）均未达到常规显著性水平；万得体系在所有层级上均未能产生显著正收益，部分分类（WIND1）甚至呈现负收益（-0.30%）。LLM 一级分类（LLM1：0.25%，T=0.42，p=0.68）由于颗粒度过粗，同行业范围过宽，领先-滞后效应未能被有效识别。流通市值加权组合（表 8）的整体回报水平有所下降，但相对排名基本保持一致：LLM2（1.03%，T=1.76，p=0.08）和 LLM3（1.16%，T=1.86，p=0.06）仍显示出统计显著的较为可观的正收益，而申万和万得体系在同等加权方式下均未能产生统计显著的正收益。

为进一步排除市场共同因子对组合收益的影响，表 9 基于 Fama-French 五因子（FF5）模型与中国四因子（CH4）模型对上述对冲组合的原始收益进行因子调整，计算风险调整后的 Alpha。

表 9　基于 FF5 与 CH4 模型的对冲投资组合 Alpha

分类	FF5 等权 Alpha	FF5 等权 T 值	FF5 加权 Alpha	FF5 加权 T 值	CH4 等权 Alpha	CH4 等权 T 值	CH4 加权 Alpha	CH4 加权 T 值
LLM1	0.002	0.370	0.003	0.460	0.006	0.822	0.008	1.019
LLM2	0.012	2.178	0.009	1.489	0.016	2.325	0.014	2.264
LLM3	0.016	2.999	0.012	1.979	0.018	2.835	0.015	2.141
SW1	0.006	1.185	0.001	0.136	0.011	1.940	0.007	1.122
SW2	0.009	1.750	0.006	0.924	0.012	1.973	0.009	1.315
SW3	0.013	1.835	0.011	1.195	0.016	2.270	0.015	1.636
WIND1	-0.004	-0.817	-0.001	-0.132	-0.001	-0.213	0.002	0.269

WIND2	0.005	1.016	0.003	0.505	0.012	1.881	0.012	1.516
WIND3	0.008	1.649	0.005	0.811	0.011	2.100	0.006	0.898
WIND4	0.006	1.095	0.005	0.788	0.009	1.646	0.006	0.976

注：该表展示了基于 Fama-French 五因子（FF5）和中国四因子（CH4）模型的对冲投资组合风险调整后的 Alpha，T 统计量经 Newey-West 滞后 6 阶调整。

风险因子调整后的 Alpha 检验结果（表 9）进一步强化了前述结论。在基于 LLM 分类构造的领先-滞后投资组合中，LLM2 的 FF5 等权 Alpha 为 0.012（T=2.178），CH4 等权 Alpha 为 0.016（T=2.325），均在 5%水平上显著；CH4 加权 Alpha 亦达到 5%显著水平（Alpha=0.014，T=2.264），仅 FF5 加权 Alpha 未通过常规显著性检验（T=1.489）。LLM3 的表现更为突出，其 FF5 等权 Alpha（0.016，T=2.999）和 CH4 等权 Alpha（0.018，T=2.835）均在 1%水平上显著，FF5 加权 Alpha（0.012，T=1.979）和 CH4 加权 Alpha（0.015，T=2.141）亦在 5%水平上显著。这表明，即便控制了市场、规模、价值、盈利、投资和流动性等系统性风险因子，基于 LLM 分类构造的领先-滞后投资组合仍能产生统计显著的超额收益，说明其所捕获的信息传播效应并非已知风险溢价的替代。相比之下，LLM1 在所有模型设定下的 Alpha 均不显著，表明过粗的 LLM 分类粒度不足以有效捕捉行业间的信息联动。

在传统行业分类体系中，申万分类（SW）的表现整体弱于 LLM 分类。SW3 在 CH4 等权下 Alpha 为 0.016（T=2.270），达到 5%显著水平，FF5 等权下 T=1.835 仅在 10%水平上边际显著；SW2 的 CH4 等权 Alpha 为 0.012（T=1.973），刚过 5%临界值，FF5 等权 T=1.750 处于 10%显著水平；SW1 仅在 CH4 等权下边际显著（T=1.940），其余设定下均不显著。值得注意的是，上述三组在流通市值加权下的 Alpha 均未通过常规显著性检验，表明申万分类下的超额收益主要集中于小市值股票，稳健性有限。

万得行业分类（WIND）的表现最弱。WIND1 在所有设定下 Alpha 均为负值或接近零，不具有经济与统计意义。WIND3 在 CH4 等权下表现相对较好（Alpha=0.011，T=2.100，5%显著），FF5 等权下边际显著（T=1.649，10%水平）；WIND2 和 WIND4 仅在 CH4 等权下达到 10%边际显著水平（T 分别为 1.881 和 1.646），其余设定下均不显著。此外，万得分类在加权下同样缺乏显著的 Alpha 表现。

综合而言，无论在 Alpha 的绝对水平还是统计显著性上，基于 LLM 分类构造的投资组合（尤其是 LLM2 和 LLM3）均系统性地优于传统申万和万得行业分类，且这一优势在等权和流通市值加权、FF5 和 CH4 两种因子模型下均保持一致，进一步验证了 LLM 在行业信息

联动识别方面的增量价值。

综合来看，投资组合构造检验表明，本研究基于大语言模型构建的 LLM 行业分类体系能够更准确地界定“同行业”的边界，从而有效捕获 A 股市场中信息在同行业公司间传播所产生的领先-滞后效应，为该分类体系的经济有效性提供了基于资产定价的直接证据。这一结论与上述行业间差异性和行业内相似性分析相互印证，共同支持了本研究分类体系的综合质量优势。

（二）Fama-MacBeth 回归

本小节通过 Fama-MacBeth（FMB）横截面回归进一步检验行业分类质量的经济含义。具体而言，本研究在每月对截面股票回报进行回归，考察同行业领先-滞后特征（MOM）能否预测个股未来一期回报，以及这一效应是否在高价股中更为显著。参照 Du 等（2025）关于高价股动量特征更为突出的理论框架，本研究预期：若某一行业分类体系能够更准确地界定"同行业"范围，则基于该分类构造的同行业回报动量与高价股哑变量的交叉项系数应当更大且更为显著——这意味着信息在更精准同行业边界内的传播效应能够更有效地被资产价格所反映。

截面回归模型的设定如下：

$$Ret_{i,t} = \alpha + \beta_1 MOM_{i,t-1} \times HighPrice_{i,t-1} + \beta_2 MOM_{i,t-1} + \beta_3 HighPrice_{i,t-1} + \gamma Controls_{i,t-1} + \epsilon_{i,t} \quad (7)$$

其中，$Ret_{i,t}$为股票 i 在 t 月的回报率，$MOM_{i,t-1}$为 t-1 月末基于各行业分类体系构造的股票 i 过去 11 个月（t-12 至 t-2 月）同行业公司等权平均回报（排除自身），$HighPrice_{i,t-1}$为高价股哑变量（收盘价位于 t-1 月市场第 90 百分位以上取 1，否则取 0）。MOM与$HighPrice$的交叉项系数β_1是本研究的核心关注系数，捕获了同行业动量效应在高价股中的差异性增强。控制变量包括资产增长率（Asset Growth）、公司规模（Size）、账面市值比（Book-to-Market）和毛利率（Gross Profitability）。除回报率外，所有连续变量均在每月按第 0.5 和第 99.5 百分位进行缩尾处理。时序均值系数的显著性检验采用 Newey-West 方法（滞后 6 阶）进行调整。

表 10 和表 11 分别报告了不含控制变量和含控制变量情形下的 FMB 回归结果，各列依次对应 LLM 三级分类（LLM1—LLM3）、申万三级分类（SW1—SW3）和万得四级分类（WIND1—WIND4）。

表 10 Fama-MacBeth 回归（不含控制变量）

	LLM1	LLM2	LLM3	SW1	SW2	SW3	WIND1	WIND2	WIND3	WIND4

$MOM \times HighPrice$	0.0085	0.0175**	0.0124	0.0067	0.0028	-0.0069	0.0085	0.0044	0.0098	0.0106
	(0.96)	(2.21)	(1.56)	(0.73)	(0.35)	(-0.23)	(0.81)	(0.52)	(1.32)	(1.34)
$HighPrice$	-0.0006	-0.0005	-0.0005	-0.0005	0.0004	-0.0015	0.0020	0.0011	-0.0002	-0.0005
	(-0.16)	(-0.13)	(-0.14)	(-0.16)	(0.12)	(-0.26)	(0.49)	(0.34)	(-0.06)	(-0.15)
MOM	0.0027	0.0047	0.0073	0.0064	0.0061	-0.0122	-0.0007	0.0062	0.0068	0.0065
	(0.30)	(0.68)	(1.13)	(0.87)	(0.86)	(-0.50)	(-0.06)	(0.80)	(0.97)	(1.07)
控制变量	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
观测值	189541	167765	144379	185929	145388	96758	191053	184996	171451	153838
平均 R^2	0.0281	0.0291	0.0315	0.0300	0.0357	0.0562	0.0294	0.0296	0.0290	0.0301

注：该表展示了同行业领先-滞后特征对个股月度收益率的预测能力。被解释变量为个股月度回报率。MOM为同行业领先-滞后特征（t-12 至 t-2 月同行业等权平均回报），$HighPrice$为高价股哑变量（收盘价高于当月市场第 90 百分位取 1）。括号内为经 Newey-West（滞后 6 阶）调整的 t 统计量。***、**、*分别代表 1%、5%、10%的显著性水平。

表 11 Fama-MacBeth 回归（含控制变量）

	LLM1	LLM2	LLM3	SW1	SW2	SW3	WIND1	WIND2	WIND3	WIND4
$MOM \times HighPrice$	0.0047	0.0148**	0.0082	0.0008	-0.0046	-0.0280	0.0050	0.0004	0.0051	0.0054
	(0.53)	(2.05)	(1.17)	(0.10)	(-0.64)	(-0.65)	(0.46)	(0.04)	(0.69)	(0.70)
$HighPrice$	0.0010	0.0010	0.0016	0.0006	0.0025	0.0036	0.0040	0.0026	0.0021	0.0019
	(0.31)	(0.34)	(0.55)	(0.22)	(0.96)	(0.57)	(1.07)	(0.93)	(0.75)	(0.77)
MOM	0.0027	0.0049	0.0072	0.0066	0.0079	-0.0055	0.0005	0.0065	0.0070	0.0064
	(0.31)	(0.81)	(1.29)	(1.03)	(1.29)	(-0.29)	(0.05)	(0.96)	(1.15)	(1.18)
控制变量	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
观测值	186366	165326	141940	183154	143788	95224	187740	182221	169432	152374
平均 R^2	0.0682	0.0721	0.0771	0.0696	0.0811	0.1119	0.0693	0.0697	0.0707	0.0742

注：同表 10。控制变量包括 Asset Growth、Size、Book-to-Market 和 Gross Profitability。括号内为经 Newey-West（滞后 6 阶）调整的 t 统计量。***、**、*分别代表 1%、5%、10%的显著性水平。

从表 10（不含控制变量）的结果来看，本研究 LLM 分类体系中，LLM2 的交叉项$MOM \times HighPrice$系数为 0.0175（t=2.21），在 5%水平上显著为正，表明基于二级 LLM 分

类构造的同行业动量在高价股中具有显著更强的预测能力；LLM1（0.0085，t=0.96）和 LLM3（0.0124，t=1.56）的交叉项系数虽方向一致但未达到统计显著水平。相比之下，申万分类三个层级的交叉项系数均小于相同颗粒度的 LLM 分类的系数，且均不显著，万得分类同样不显著。

表 11 加入资产增长率、公司规模、账面市值比和毛利率四个控制变量后，结果呈现出较高的稳健性。LLM2 的交叉项系数为 0.0148（t=2.05），维持在 5%显著性水平，说明其对高价股同行业动量效应的捕获能力并非来源于对已知风险因子的被动暴露。其他分类体系的交叉项系数在加入控制变量后均统计不显著，与不含控制变量的结果高度一致。

综合来看，FMB 回归检验表明，在同行业领先-滞后动量与高价股的交互效应这一维度上，本研究构建的 LLM 二级分类（LLM2）具有最强的预测能力。这一结论与投资组合构造检验中 LLM2 和 LLM3 表现突出的发现相互印证——LLM 行业边界更为精准，信息传播的时序特征得以更清晰地体现在高价股的回报预测中，而申万和万得体系的行业划分则不足以有效识别这一效应。

七、结论

本文立足于当前中国上市公司行业分类更新滞后与区分度不足等核心痛点，使用上市公司年报文本，引入大语言模型的文本嵌入技术与聚类算法，构建了一套完全由数据驱动、涵盖三级行业的中国上市公司动态行业分类体系。在此基础上，本文从行业间差异性、行业内相似性、投资组合收益及横截面定价等多个维度，系统评估了该分类体系相较于主流标准的分类准确性。

本文研究发现，所构建的 LLM 分类体系在分类质量核心维度上显著优于同颗粒度的申万、万得及中上协分类，其在多项财务指标上的行业间标准差更高、行业固定效应回归 R^2 更高，能更好实现“类内相似、类间差异”的分类标准，且结论经稳健性检验依然成立；资产定价领域的拓展分析进一步验证了该体系的有效性；同时，本文在方法论层面突破了传统行业分类范式，依托文本嵌入模型、自底向上的嵌套聚类流程及大语言模型的两阶段命名策略，打造出兼具层级连贯性、数据客观性与语义可解释性的分类体系，为金融文本分析应用拓展和行业分类动态更新奠定了方法论基础。

综上所述，本文构建的中国 A 股上市公司行业分类数据集在各项分类指标上均展现出显著优势。这不仅为金融经济学实证研究提供了一套更精确、更动态的基础分析工具，也为

监管层优化行业统计标准、为市场参与者深化产业认知提供了有益参考。未来研究可进一步引入多源文本信息、探索更先进的嵌入模型与聚类算法，或将该分类体系应用于公司估值、同业比较、产业链分析等更广泛的研究场景，以持续推动中国资本市场基础数据建设与学术研究的高质量发展。

参考文献：

（1）段丙蕾、汪荣飞、张然：《南橘北枳：A 股市场的经济关联与股票回报》，《金融研究》，2022 年第 2 期。

（2）郭鹏飞、孙培源：《资本结构的行业特征：基于中国上市公司的实证研究》，《经济研究》，2003 年第 5 期。

（3）洪永淼、汪寿阳：《大数据如何改变经济学研究范式？》，《管理世界》，2021 年第 10 期。

（4）洪永淼、汪寿阳：《ChatGPT 与大模型将对经济学研究范式产生什么影响？》，《计量经济学报》，2024 年第 1 期。

（5）胡楠、邱芳娟、梁鹏：《竞争战略与盈余质量——基于文本分析的实证研究》，《当代财经》，2020 年第 9 期。

（6）姜富伟、孟令超、唐国豪：《媒体文本情绪与股票回报预测》，《经济学（季刊）》，2021 年第 4 期。

（7）陆瑶、施函青、周欣怡：《中国企业数字技术风险暴露对企业价值的影响——来自大语言模型的文本分析证据》，《经济研究》，2025 年第 2 期。

（8）沈艳、陈赟、黄卓：《文本大数据分析在经济学和金融学中的应用：一个文献综述》，《经济学（季刊）》，2019 年第 4 期。

（9）徐巍、梁上坤、钱宇航：《资本市场出清与实体经济投资——基于上市公司强制退市的实证研究》，《管理世界》，2025 年第 8 期。

（10）Alford，A. W.，1992，"The Effect of the Set of Comparable Firms on the Accuracy of the Price-earnings Valuation Method"，*Journal of Accounting Research*，30(1)，pp.94~108.

（11）Bhojraj，S.，Lee，C. M. and Oler，D. K.，2003，"What's My Line? A Comparison of Industry Classification Schemes for Capital Market Research"，*Journal of Accounting Research*，41(5)，pp.745~774.

（12）Breitung，C. and Müller，S.，2025，"Global Business Networks"，*Journal of Financial Economics*，166，p.104007.

（13）Clarke，R. N.，1989，"SICs as Delineators of Economic Markets"，*Journal of*

Business，pp.17~31.

（14）Devlin，J.，Chang，M. W.，Lee，K. and Toutanova，K.，2019，"Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding"，In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*，pp.4171~4186.

（15）Du，J.，Huang，D.，Liu，Y. J.，Shi，Y.，Subrahmanyam，A. and Zhang，H.，2025，"Nominal Prices，Retail Investor Participation，and Return Momentum"，*Management Science*，72(3)，pp.2064~2089.

（16）Frank，M. Z. and Goyal，V. K.，2009，"Capital Structure Decisions: Which Factors Are Reliably Important?"，*Financial Management*，38(1)，pp.1~37.

（17）Hoberg，G. and Phillips，G.，2010，"Product Market Synergies and Competition in Mergers and Acquisitions: A Text-based Analysis"，*The Review of Financial Studies*，23(10)，pp.3773~3811.

（18）Hoberg，G. and Phillips，G.，2016，"Text-based Network Industries and Endogenous Product Differentiation"，*Journal of Political Economy*，124(5)，pp.1423~1465.

（19）Hoberg，G. and Phillips，G. M.，2025，"Scope，Scale，and Concentration: The 21st-century Firm"，*The Journal of Finance*，80(1)，pp.415~466.

（20）Kahle，K. M. and Walkling，R. A.，1996，"The Impact of Industry Classifications on Financial Research"，*Journal of Financial and Quantitative Analysis*，31(3)，pp.309~335.

（21）Le，Q. and Mikolov，T.，2014，"Distributed Representations of Sentences and Documents"，*International Conference on Machine Learning*，pp.1188~1196.

（22）Li，K.，Mai，F.，Shen，R.，Yang，C. and Zhang，T.，2026，"Dissecting corporate culture using generative AI"，*The Review of Financial Studies*，39(1)，pp.253~296.

（23）Loughran，T. and McDonald，B.，2011，"When Is a Liability Not a Liability? Textual Analysis，Dictionaries，and 10-Ks"，*The Journal of Finance*，66(1)，pp.35~65.

（24）McGahan，A. M. and Porter，M. E.，1997，"How Much Does Industry

Matter，Really?"，*Strategic Management Journal*，18(S1)，pp.15~30.

（25）Mikolov，T.，Chen，K.，Corrado，G. and Dean，J.，2013，"Efficient Estimation of Word Representations in Vector Space"，*arXiv preprint arXiv:1301.3781*.

（26）Moskowitz，T. J. and Grinblatt，M.，1999，"Do Industries Explain Momentum?"，*The Journal of Finance*，54(4)，pp.1249~1290.

（27）Siano，F.，2025，"The News in Earnings Announcement Disclosures: Capturing Word Context Using LLM Methods"，*Management Science*，71(11)，pp.9831~9855.

（28）Tetlock，P. C.，2007，"Giving Content to Investor Sentiment: The Role of Media in the Stock Market"，*The Journal of Finance*，62(3)，pp.1139~1168.

附录：大语言模型提示词（Prompts）

本附录说明了行业分类体系命名流程，并列出了行业分类体系命名过程中，各阶段使用的大语言模型提示词，其中花括号{}内的内容为动态替换的变量。为了保持输出的稳定性，大模型的温度参数被设置为 0.

一、一级行业分类

（一）一级行业业务画像总结

说明：针对每个一级行业聚类，将该聚类下所有公司的 MDA 文本片段输入模型，生成行业业务画像总结。使用模型为 Qwen-Long。每篇 MDA 截取前 1000 字，每个聚类最多抽取 1500 家公司。以下是完整的提示词：

你是一名专业的行业业务分析师。我输入了一批中国 A 股上市公司“管理层讨论与分析”的部分文本，这些公司由聚类算法汇总为同一个一级行业，请你撰写一份非常详尽、全面的行业业务画像总结。请基于提供的该聚类下所有公司的全量文本进行深度分析，确保在不遗漏任何主要业务形态的前提下，详尽描述并且突出展示该群体的核心业务特征（涵盖核心产品、关键技术及服务对象），并对聚类内部可能存在的细分业务方向进行逐一梳理与列示，最终输出一份篇幅不超过 6000 字的详实业务总结。请直接输出总结内容。

以下是聚类 ID {聚类 ID} 包含的 MDA 文本片段 (共 {N} 篇)：

文档 {股票代码}_{年份}: {MDA 文本前 1000 字}

文档 {股票代码}_{年份}: {MDA 文本前 1000 字}

.....

（二）一级行业全局命名

说明：将所有一级行业的 ID 与业务总结一次性输入模型，进行全局对比后统一命名。使用模型为 Qwen3-Max，启用 JSON 输出模式。以下是完整的提示词：

你是一位资深的金融行业分类专家。你需要接收用户提供的多组“行业 ID”及其“业务总

结”，通过对比分析，为每一个行业赋予一个一级行业名称。你必须详细对比所有行业 ID 的业务总结，再赋予类别名称，不同 ID 的名称必须有明显区分，绝对禁止重复。名称必须精准覆盖该类下绝大多数公司的核心业务。在命名规则上使用中国 A 股市场通用的行业术语（如：基础化工、公用事业、食品饮料、高端装备、电子元件等）。类别名称长度严格控制在 2-6 个中文字符，不要包含标点符号。请仅返回一个标准的 JSON 对象，JSON 的 Key 必须是输入的行业 ID，Value 是你命名的行业名称，不要包含任何解释性文字。

输出示例： {

"1": "石油化工",

"2": "医药生物",

"15": "食品饮料" }

待命名数据列表：

ID: {聚类 ID_1} Content: {该聚类的业务画像总结全文}

ID: {聚类 ID_2} Content: {该聚类的业务画像总结全文}

.....

二、二级行业分类

（一）二级行业业务画像总结

说明：针对每个二级行业聚类，同时提供其所属一级行业的名称与总结作为背景信息，连同该二级聚类下所有公司的 MDA 文本一起输入模型。使用模型为 Qwen-Long。以下是完整的提示词：

你是一名专业的行业业务分析师。我输入了一批中国 A 股上市公司“管理层讨论与分析”的部分文本，这些公司由聚类算法汇总为同一个二级行业，请你撰写一份非常详尽、全面的行业业务画像总结。为了帮助你分析，我还输入了这些公司所属的一级行业的名称和行业业务总结。请基于提供的一级行业的名称和行业业务总结以及该二级聚类下所有公司的全量文本进行深度分析，确保在不遗漏任何主要业务形态的前提下，详尽描述并且突出展示该群体的

核心业务特征（涵盖核心产品、关键技术及服务对象），并对聚类内部可能存在的细分业务方向进行逐一梳理与列示，最终输出一份篇幅不超过 6000 字的详实业务总结。请直接输出总结内容。

【参考：该二级行业所属的一级行业信息】

一级行业名称：{一级行业名称}

一级行业总结：{一级行业业务画像总结}

【核心分析数据：二级行业包含的所有 MDA 文本片段】

以下是该二级行业（ID: {二级聚类 ID}）包含的所有 MDA 文本片段：

文档 {股票代码}_{年份}: {MDA 文本前 1000 字}

文档 {股票代码}_{年份}: {MDA 文本前 1000 字}

.....

（二）二级行业全局命名

说明：按一级行业分组调用，每次将该一级行业的背景信息及其下属所有二级行业的全量业务画像输入模型。使用模型为 Qwen3-Max，启用 JSON 输出模式。该阶段的温度参数设为 0.1。以下是完整的提示词：

你是一位资深的金融行业分类专家。请基于一级行业【{一级行业名称}】的背景，深度分析其下属若干二级子行业的全量业务描述，为每个二级子行业赋予一个精准的名称。名称必须在语境上体现出对【{一级行业名称}】的从属或细分关系。即使业务存在跨界，命名时也应侧重描述其在该一级行业视角下的特定属性（例如：一级为“汽车”，二级应命名为“汽车零部件”而非单纯的“机械加工”）。在我的分类标准中还有三级行业，因此你也不能够写的过于详细或者具体，颗粒度应该介于一级行业和三级行业之间。你必须详细对比所有二级行业 ID 的业务总结，再赋予类别名称，不同 ID 的名称必须有明显区分，绝对禁止重复。名称必须精准覆盖该类下绝大多数公司的核心业务。二级分类名称应使用中国 A 股市场通用的二级行业术语，严格控制二级行业名称长度为 4-10 个中文字符，严禁包含标点。请仅返回

一个标准的 JSON 对象，JSON 的 Key 必须是输入的二级行业 ID，Value 是你命名的行业名称，不要包含任何解释性文字。

输出示例： {

"19_1": "化学原料药",

"19_2": "生物制品与疫苗",

"19_3": "医疗器械耗材" }

【一级行业背景】

ID: {一级行业 ID}

名称: {一级行业名称}

总结: {一级行业业务画像总结}

【待命名的二级子行业列表】

以下是属于该一级行业的所有子集，请根据其详细描述进行命名：

二级行业 ID: {二级聚类 ID_1}

二级行业全量业务画像: {该二级行业的业务画像总结全文}

二级行业 ID: {二级聚类 ID_2}

二级行业全量业务画像: {该二级行业的业务画像总结全文}

.....

【输出要求】

请仅返回一个 JSON 对象，格式如 System Prompt 所示。