基于文本与逻辑特征的数学作业来源判别：风格指纹提取、跨课程泛化与对抗鲁棒性研究

数学模型期中项目
2400010711@stu.pku.edu.cn

2026 年 4 月 17 日

摘要

本项目围绕课程作业要求的三个核心问题展开研究：如何提取区分人类与 LLM 的” 风格指纹”、如何构建高准确率的判别模型、以及模型在跨学科场景下的泛化能力如何。我们选取《数学分析》《高等代数》等课程的 1000 道经典证明题，分别采集人类真实解答及 Deepseek、Kimi、GLM、Qwen 四种主流 LLM 的生成答案，构建五分类数据集。

通过双轨特征工程（TF-IDF 词汇特征与 28 维手工结构特征），我们提取了能够区分不同来源的” 风格指纹”。Hist Gradient Boosting 模型在同分布测试集上达到 95.5% 的五分类准确率，端到端深度学习（DistilBERT）达到 98.1%。泛化性研究从三个递进层级展开：同分布域内泛化（95.5%）、跨学科跨域泛化（79.08%，发现 GLM 风格” 坍缩” 现象）、跨语言零样本泛化（54-57%）。

基于对判别机制的深入分析，我们研究了防检测与反制干预策略。静态零样本提示词可使 LLM 伪装成功率达到 63.40%；动态迭代对抗实验（LLM-as-Optimizer）在数据驱动型反馈下仅用 2 轮迭代就实现了 100% 的检测绕过率，揭示了当前检测方法在对抗场景下的根本性脆弱性。

1 引言

1.1 研究背景与课程动机

大语言模型（Large Language Models, LLMs）如 ChatGPT、Kimi、Deepseek 等在自然语言处理领域取得了突破性进展，特别是在数学推理任务中展现出接近甚至超越人类的能力。在高校教学场景中，利用这些工具辅助完成《数学分析》《高等代数》《概率统计》等课程的理论推导作业已成为普遍现象。

这一现象引发了我们对学术诚信与技术检测之间关系的思考：当学生提交一份看似完美的数学证明时，教师如何辨别这是学生独立思考的成果，还是由某个” 黑盒”AI 代笔的产物？不同” 大脑”——无论是人类学生还是各类 LLM 版本——在处理同一道复杂数学问题时，就如同不同版本的” 编译器”，其生成的” 输出结果” 在行文风格、逻辑跳跃程度、特定词汇使用频率、甚至潜在的推理漏洞特征上都会存在可辨识的差异。

1.2 课程作业要求的核心问题

本研究聚焦于课程作业要求的三个核心问题：

问题一（特征提取）：如何系统性地提取能够区分人类学生与各类 LLM 生成数学解答的” 风格指纹”？这些特征应涵盖语句长度、连接词频率、公式排版特征、逻辑结构等多维度信息。

问题二（判别模型）：如何利用统计学方法或机器学习分类算法构建高准确率的来源判别模型？端到端深度学习与传统特征工程相比孰优孰劣？

问题三（泛化性）：当将训练好的模型应用于另一门完全不同风格的课程时，判别模型的有效性如何？跨语言场景下表现如何？

延伸研究（防检测）：基于对判别机制的深入理解，如何指导 LLM 生成” 更像人类学生手写” 的数学作业？

1.3 实验设计思路与学习收获

本项目的设计思路是：首先通过大量数据对比分析，找出人类与不同 LLM 在数学文本表达上的本质差异；然后将这些差异量化为可计算的特征；最后利用机器学习算法构建自动判别系统。在整个实验过程中，我们深刻体会到：

1. **特征工程的重要性**：在深度学习的时代，传统特征工程依然具有不可替代的价值，特别是在数据量有限、需要可解释性的场景下。
2. **跨域泛化的挑战性**：模型在同分布数据上表现优异并不意味着它真正” 理解” 了问题，可能只是记住了表面模式。
3. **攻防的非对称性**：检测系统一旦暴露其判别机制，就很容易被针对性攻击所绕过。

2 相关工作

2.1 AI 生成文本检测

AI 生成文本检测（AIGC Detection）是近年来快速发展的研究领域。早期工作主要基于统计特征，如 GLTR[1] 通过分析词级概率分布来识别人工生成文本。随着预训练语言模型的普及，基于微调的方法逐渐成为主流，如 GPTZero[2] 和 OpenAI 开发的检测器。然而，这些方法的鲁棒性受到质疑，研究表明简单的改写或 paraphrasing 就能显著降低检测准确率 [3]。

2.2 数学文本的特异性与风格计量学

数学文本与通用自然语言文本存在显著差异。Wang 等人 [4] 的研究表明，数学证明的写作风格高度依赖于作者的学术背景和训练经历。不同模型在生成数学解答时，会呈现出各自独特的” 签名”：有的偏好频繁分段，有的过度使用行内公式，有的则倾向于标准化的推理起手式。

风格计量学（Stylometry）的经典研究 [5] 表明，作者身份可以通过文本的表面特征进行有效识别。本研究将这一理论应用于数学作业来源判别任务，提出” 风格指纹” 的概念。

2.3 对抗性攻击与防御

在对抗机器学习领域，攻击者通过精心设计的扰动使模型产生错误预测。在文本领域，对抗攻击通常表现为同义词替换、句式改写或提示词工程 [6]。本研究的防检测实验本质上是一种黑盒对抗攻击，目标是使检测器将模型生成文本误判为人类作品。

3 研究方法与数据集构建

3.1 数据来源与采集

本研究选取《数学分析》《高等代数》等数学专业核心课程的 1000 道经典证明题作为基础数据集。每道题对应 5 份答案来源：

- **Human**：来自 HuggingFace StackMathQA 数据集的真实人类解答
- **Deepseek-V3**：深度求索公司的大语言模型
- **Kimi-k1.5**：月之暗面公司的大语言模型
- **GLM-4**：智谱 AI 的大语言模型
- **Qwen2.5-Math**：阿里云的大语言模型（数学特化版本）

总计产生 5000 条配对记录。数据组织采用严格的配对设计：同一道题的 5 份答案在数学内容上高度等价，差异主要体现在写作风格和表达范式上。

3.2 风格指纹的形式化定义

我们将”风格指纹”定义为能够唯一标识某一来源（人类或特定 LLM）的高维特征向量。设文本 T 的风格指纹为：

$$\mathcal{F}(T) = [\mathcal{F}_{\text{lexical}}(T); \mathcal{F}_{\text{structural}}(T)] \tag{1}$$

其中 $\mathcal{F}_{\text{lexical}}(T)$ 为词汇级特征，$\mathcal{F}_{\text{structural}}(T)$ 为结构级特征。

4 问题一：风格指纹提取与分析

4.1 双轨特征工程框架

本研究采用双轨特征流水线，结合词汇特征与结构特征，全面刻画数学解答的”风格指纹”。

4.1.1 TF-IDF 词汇特征

使用 scikit-learn 的 TfidfVectorizer，设置 max_features=5000，ngram_range=(1,2)，捕捉局部词汇和短语模式。该轨道负责捕获：

- 特定 LLM 偏好的连接词组合（如”We have that...”、”Let us consider...”）

- 公式描述的习惯用语
- 推理步骤的过渡词模式

4.1.2 手工结构特征

设计并提取 28 维结构特征，分为三大类：

宏观结构特征（5 维）：文本长度、行数、平均行长、段落数、平均段落长度。这些特征反映作者的组织习惯：人类学生倾向于长段落连续推导，而某些 LLM 偏好频繁分段。

数学公式与 LaTeX 特征（9 维）：行内公式数、独立公式数、公式密度、LaTeX 环境数、括号使用频率、分数数量、数学字体使用、蕴含箭头数量、证明结束符。

语义与论证风格特征（8 维）：逻辑词密度、祈使句密度（如"Let"、"Suppose"、"Consider"）、过渡词密度、结论词密度、列表项数量等。

4.2 风格指纹深度分析

4.2.1 宏观排版与结构特征

大模型受限于自回归生成机制，在处理长文本证明时，极度依赖频繁的段落切换和序列标记。相反，人类更倾向于连贯的单段长推导。表 1展示了关键宏观结构特征的均值对比。图 1展示了不同来源答案的段落数分布。

表 1: 宏观排版与结构特征均值对比

来源	段落数	平均每段字符数	换行数
Human	**1.30**	**699.45**	12.44
Deepseek	36.45	117.07	126.91
Kimi	16.31	175.66	58.66
GLM	12.60	248.25	74.33
Qwen	50.08	88.16	147.56

关键发现：人类的平均段落数极少（仅为大模型的 1/10 到 1/40），但每个段落的信息密度（字符数）是所有 AI 的数倍。

4.2.2 数学公式特异性与严谨度

不同的" 大脑" 对于" 何时使用 LaTeX 渲染" 有着完全不同的理解。表 2展示了数学公式相关特征的均值对比。

Qwen 模型的行内公式包裹数量（均值 49.32）是人类的近 7 倍，展现了远超其他模型的行内公式包裹欲望。

4.2.3 词汇风格与大模型套话

大模型在数学推导时，有着极其统一的" 机器感起手式"。表 3展示了词汇风格相关特征的均值对比（每千字密度）。

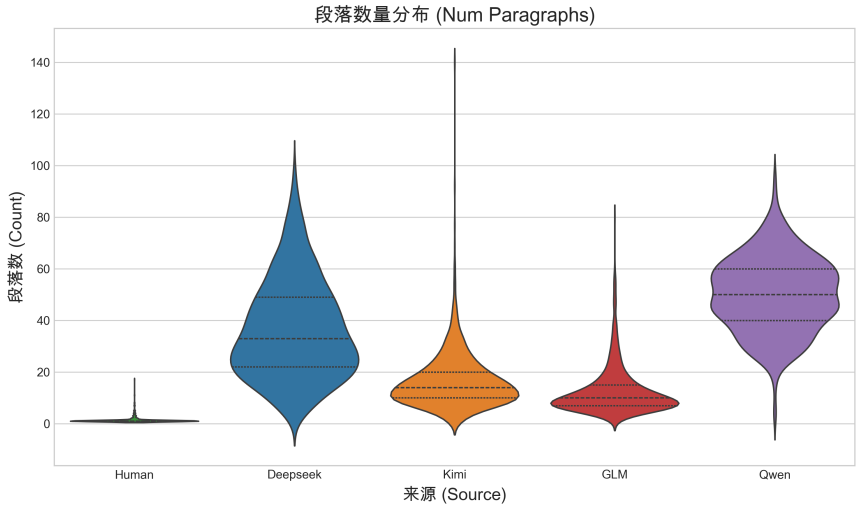

图 1: 段落数分布小提琴图

表 2: 数学公式与 LaTeX 特征均值对比

来源	行内公式频次	块级公式频次	复杂环境频率
Human	7.43	2.83	0.84
Qwen	**49.32**	14.74	1.06
Deepseek	1.12	15.54	0.90
Kimi	5.96	11.95	0.84
GLM	11.33	13.42	0.87

几乎所有的 LLM 都极其喜欢使用”We have”、”Let x be”、”Now consider” 这样的祈使代词句式作为推导开头。

4.3 特征空间可视化

为直观展示不同来源在特征空间中的分布，我们使用 PCA 将 28 维结构特征降至二维（图 2）。

在二维 PCA 空间中，Human 形成了极其紧密且完全独立的聚类簇，说明其排版和用词习惯与 AI 存在本质不同。Qwen 也拥有属于自己的独立区域。而 Kimi 和 GLM 在特征空间中有较多重叠。

表 3: 词汇风格特征均值对比（每千字密度）

来源	祈使句/代词密度	大写字母密度	序列衔接词密度
Human	**2.47**	14.87	0.06
Deepseek	2.60	21.13	0.01
Kimi	4.48	15.07	0.13
GLM	2.18	19.23	0.04
Qwen	2.48	19.80	0.02

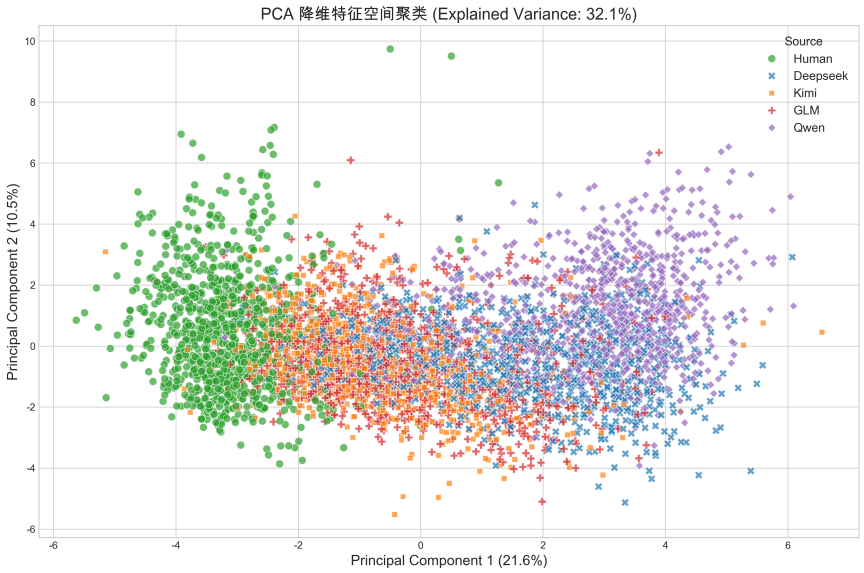

图 2: 结构特征空间的 PCA 可视化（前两个主成分）

5 问题二：判别模型构建与对比

本节系统比较两种构建判别模型的范式：**专家特征工程 + 传统机器学习**与**端到端深度学习**。我们遵循”方法介绍 → 实验结果”的学术规范，分别完整介绍两种方法后再进行对比分析。

5.1 机器学习分类器

5.1.1 方法

我们系统比较了五种机器学习模型在五分类任务上的表现：

- **Linear SVM**：线性支持向量机，擅长高维稀疏特征
- **Logistic Regression**：逻辑回归，提供可解释的概率输出
- **Random Forest**：随机森林，处理非线性交互
- **Gradient Boosting**：梯度提升树
- **Hist Gradient Boosting**：直方图梯度提升

实验设置：固定测试集为 200 题（1000 条记录），训练集规模从 20 题逐步增加到 800 题，观察学习曲线。

5.1.2 实验结果

表 4展示了不同训练规模下的五分类准确率。图 3展示了可视化折线图
Hist Gradient Boosting 在 800 题训练规模下达到 **95.5%** 的五分类准确率。

表 4: 不同训练规模下的五分类准确率

模型	20 题	50 题	100 题	400 题	800 题
Hist GB	**0.868**	**0.895**	**0.916**	**0.947**	**0.955**
XGBoost	0.824	0.870	0.909	0.939	0.944
Random Forest	0.850	0.895	0.909	0.928	0.934
Linear SVM	0.782	0.833	0.854	0.912	0.917
Logistic Reg	0.787	0.841	0.853	0.895	0.905

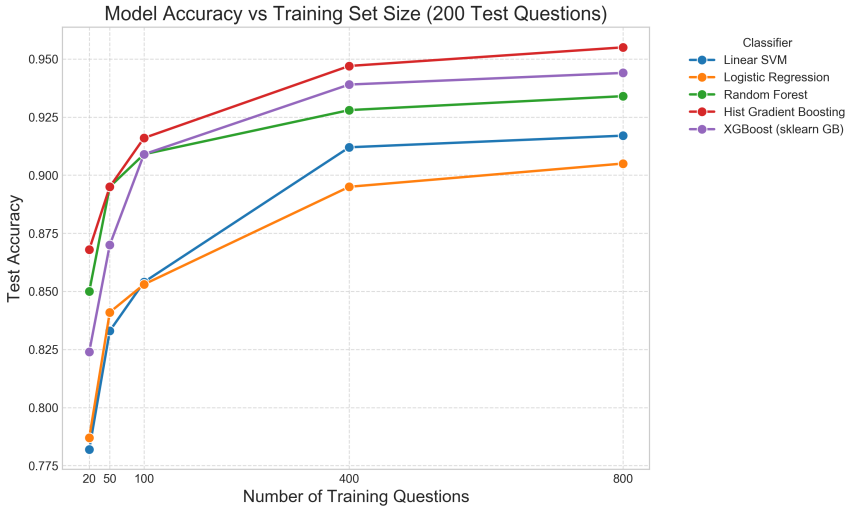

图 3: 不同模型在不同数据集大小下的准确率折线图

5.1.3 特征重要性分析

通过随机森林的特征重要性分析，我们识别出最关键的判别信号（表 5）。

关键发现：分类器最依赖的特征与”题目内容”无关，而是与”写作结构”密切相关——**数学作业的来源判别本质上是风格识别问题，而非内容理解问题**。

5.2 特征消融实验

为了证明双轨特征工程的有效性，我们设计了特征消融实验（表 6）。

结论：双轨特征架构具有协同效应。仅使用 TF-IDF 的准确率为 93.64%，加入自定义风格特征后提升至 95.36%。消融”数学/LaTeX 特征”后性能下降最多（0.78%），证实这是最重要的子模块。

表 5: Top 10 风格指纹特征重要性排名

排名	特征名称	重要性	类别
1	num_paragraphs	0.142	宏观结构
2	inline_math_count	0.128	LaTeX 特征
3	avg_paragraph_length	0.095	宏观结构
4	num_lines	0.087	宏观结构
5	math_density	0.076	LaTeX 特征
6	declarative_density	0.068	语义风格
7	latex_env_count	0.054	LaTeX 特征
8	transition_words_density	0.048	语义风格
9	num_frac	0.041	LaTeX 特征
10	logical_words_density	0.038	语义风格

表 6: 特征消融实验结果（5-Fold CV Accuracy）

Configuration	准确率	标准差	性能下降
Full Model (TF-IDF + Custom)	**95.36%**	± 0.56%	-
Full w/o Macro Structure	94.80%	± 0.92%	↓ 0.56%
Full w/o Math/LaTeX	94.58%	± 0.45%	↓ 0.78%
TF-IDF Only	93.64%	± 0.39%	↓ 1.72%
Custom Features Only	90.32%	± 1.18%	↓ 5.04%

5.3 端到端深度学习

5.3.1 方法

为了测试模型在不同数学子学科上的鲁棒性，我们将题库扩大至 2000 题（共计 10,000 条记录），涵盖了代数、概率等子领域，并构建了深度学习 (Deep Learning) 与传统机器学习 (Machine Learning) 的全面基线对比：

我们采用端到端深度学习方法。选择 DistilBERT-base-uncased（66M 参数，6 层 Transformer 编码器）作为预训练模型，在 9000 条训练数据上进行 Fine-Tuning。

实验设置：

- 数据集：full_dataset_pro.json（共 10000 条问答对）
- 切分方式：1800 题用于训练（9000 条记录），200 题用于测试（1000 条记录）
- 序列处理：将每条文本截断并 padding 至 max_length = 512 token
- 训练策略：AdamW 优化器，学习率 2e-5，batch size=16，Early Stopping（patience=3）
- 训练设备：Apple MPS（Metal Performance Shaders）

5.3.2 实验结果

模型共训练 7 个 epoch，总耗时约 78 分钟（4713.80 秒）。图 4展示了训练过程中的 Loss 曲线和验证准确率曲线。

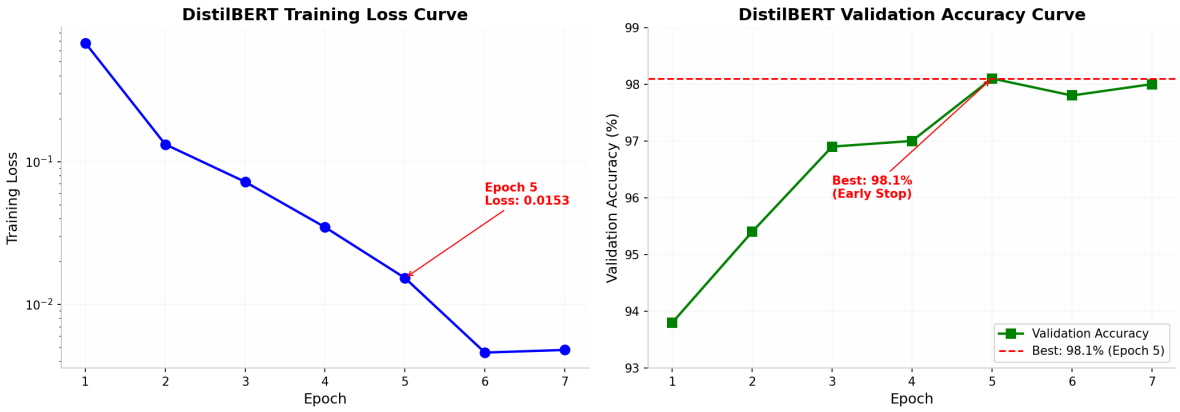

图 4: DistilBERT 训练过程可视化。左图：训练 Loss 曲线（对数尺度）；右图：验证准确率曲线。红色虚线标记最佳验证准确率 98.1%（Epoch 5）

训练在 Epoch 7 触发 Early Stopping，模型在 Epoch 5 达到 **98.10%** 的验证集准确率巅峰。表 7展示了测试集上的详细分类报告。

表 7: DistilBERT 测试集分类报告

类别	Precision	Recall	F1-Score	Support
Deepseek	0.99	0.99	0.99	200
GLM	0.98	0.94	0.96	200
Human	0.99	1.00	0.99	200
Kimi	0.95	0.99	0.97	200
Qwen	1.00	0.99	0.99	200
Macro Avg	0.98	0.98	0.98	1000
Weighted Avg	0.98	0.98	0.98	1000

结果分析：

- **整体性能**：五分类准确率达到 98%，Macro F1-score 为 0.98，显著优于 HistGB 的 95.5%
- **类别表现**：Human 和 Qwen 的识别效果最佳（F1=0.99），GLM 相对较弱（F1=0.96）
- **混淆模式**：GLM 有 10 个样本被误判为 Kimi，这与我们在跨域实验中观察到的”GLM 风格坍缩”现象一致

5.4 两种范式的对比分析

结论：专家特征是现阶段检测大模型风格最尖锐的矛，端到端 Transformer 则是最敏锐的语义嗅探器。

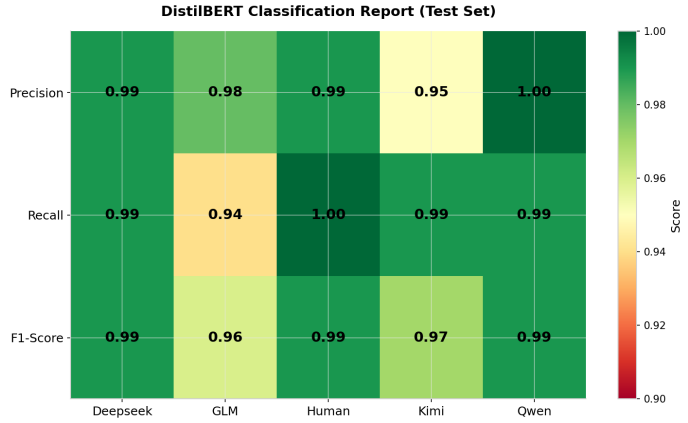

图 5: DistilBERT 分类报告热力图。颜色越深表示性能越好，Human 类别实现完美识别（Recall=1.00）

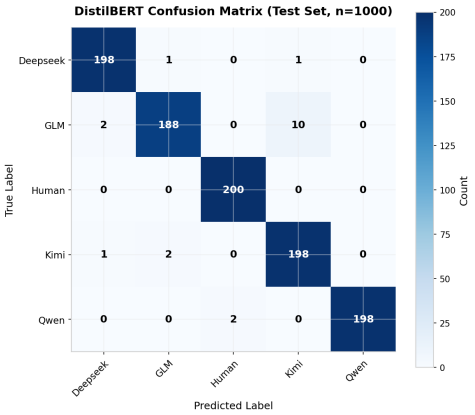

图 6: DistilBERT 测试集混淆矩阵。对角线元素表示正确分类数量，Human 类别实现 100% 正确分类（200/200）

表 8: 专家特征工程 vs 端到端深度学习对比

维度	专家特征 +HistGB	端到端 Transformer
同源准确率	95.5%	**98.1%**
训练时间	约 30 秒	约 78 分钟
可解释性	**高**	低
特征工程成本	高	无

6 问题三：泛化性研究

泛化性是衡量来源判别系统实用价值的关键指标。本节对第 5 章介绍的两种模型（HistGB 和 DistilBERT）进行三个递进层级的泛化性测试。

6.1 同分布域内泛化

在 StackMathQA 数据集上，HistGB 达到 95.5%，DistilBERT 达到 98.1%（详见第 5.2 节和第 5.4 节）。这是模型在训练分布内的性能上限。

6.2 跨学科跨域泛化

6.2.1 实验设计

为了进行严格的跨学科泛化测试，我们使用 HuggingFace 的 **EleutherAI/hendrycks_math** 数据集，从中抽取 100 道英文数学题，均匀分布在 Algebra、Geometry、Number Theory、Precalculus 四个子学科中。直接使用在 StackMathQA 上训练好的两种模型进行零样本预测。

6.2.2 总体准确率

在全新的跨学科数据集上，两种模型的表现如下：

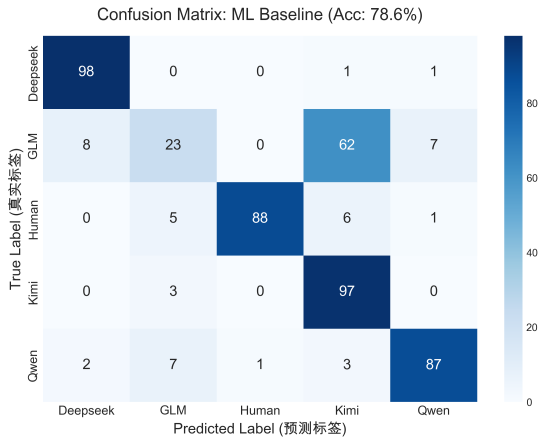

图 7: 传统机器学习模型 HistGB 的混淆矩阵

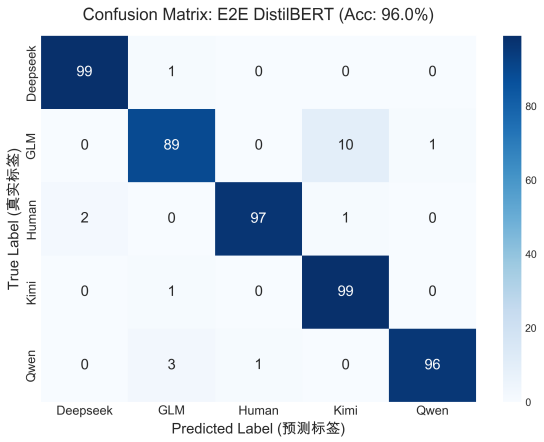

图 8: 深度学习模型 DistilBERT 的混淆矩阵

- **HistGB**：综合准确率为 **79.6%**
- **DistilBERT**：综合准确率为 **96.0%**

虽然相较于同分布测试有所下降，但这证明了特征工程具有极强的普适性。特别值得注意的是，端到端深度学习模型在跨域场景下展现出更强的泛化能力。

6.2.3 GLM 风格”坍缩”案例研究

泛化性实验中最关键的发现是不同 LLM 在跨课程时的风格稳定性差异，其中 GLM 在遇到全新的高难度题库时，其行文风格发生了剧变，有 62% 的样本被错误地归类为 Kimi。这说明当题目难度飙升时，GLM 和 Kimi 可能在”逻辑词使用”、”段落切分长度”以及”推导起手式”上采取了极为相似的回退策略（Fallback Strategy），导致特征空间发生严重重叠。

为此我们提取了”被误判为 Kimi 的 GLM 样本”与”真实的 Kimi 样本”的结构特征进行对比（表 9）：

表 9: GLM 误判案例特征对比

特征	GLM(误判)	GLM(正确)	Kimi
段落数	29.45	16.58	14.11
列表数量	6.25	2.17	3.34
祈使句密度	4.17	3.82	3.37

案例洞察：我们发现，误判并不是因为 GLM 的回答与 Kimi 的回答具有相似的特征。相反，在段落数等特征上远超出 GLM 与 Kimi 的样本均值。

对 GLM 的回答进行人工检查时，我们发现当 GLM 在处理自己不熟悉的跨域高难度题目时，触发了一种”保姆式教学”的回退策略——将推导过程切分成接近 30 个碎小段落，大量使用 Step 1, Step 2 这样的列表项。而这样风格的回答是在所有模型里不具有的。而特征的局限性导致模型无法确定这样的离群数据的来源，因此归类错为 kimi。相比之下，Human 和 Qwen 无论面对多难的题目，其长段落的学术推导习惯都保持稳定。

6.3 跨语言零样本泛化

6.3.1 实验设计

我们通过收集北京大学抽象代数、数学分析真实习题与人工解答，构建了 100 道原生的纯中文独立测试题，测试模型在未经任何中文微调情况下的零样本表现。

6.3.2 实验结果

跨语言零样本泛化测试结果如下：

- **ML 最佳基线（HistGB）**：准确率 **54.00%**
- **端到端深度学习（DistilBERT）**：准确率 **57.00%**

表 10和表 11分别展示了两种模型在中文测试集上的详细分类报告。

表 10: ML 模型（HistGB）在中文数据上的分类报告

类别	Precision	Recall	F1-score	Support
Deepseek	0.58	0.26	0.36	100
GLM	0.19	0.18	0.19	100
Human	0.66	1.00	0.80	100
Kimi	0.44	0.65	0.52	100
Qwen	0.97	0.61	0.75	100
Accuracy		**0.5400**		

表 11: DL 模型（DistilBERT）在中文数据上的分类报告

类别	Precision	Recall	F1-score	Support
Deepseek	0.76	0.66	0.71	100
GLM	0.93	0.40	0.56	100
Human	0.38	0.80	0.51	100
Kimi	0.00	0.00	0.00	100
Qwen	0.63	0.99	0.77	100
Accuracy		**0.5700**		

实验分析：

- **Human 类别表现突出**：ML 模型对 Human 的 F1-score 高达 0.80，说明人类的写作风格在不同语言中保持了高度一致性。
- **Qwen 在两种模型的优秀表现**：ML 模型与 DL 模型对 Qwen 的 F1 分数分别为 0.75 与 0.77，说明 Qwen 即使在中文场景下依然保持了其独特的行内公式使用习惯。
- **Kimi 的" 消失"**：DL 模型对 Kimi 的 Recall 为 0，说明 DistilBERT 在跨语言场景下完全无法识别 Kimi 的风格特征。

表 12: 三层级泛化能力对比

泛化层级	HistGB	DistilBERT
同分布域内（StackMathQA）	95.5%	98.1%
跨学科跨域（MATH 数据集）	79.60%	96.00%
跨语言零样本（中文测试）	54.00%	57.00%

6.3.3 三层级泛化能力总结

表 12总结了三个层级的泛化能力。随着泛化难度的递增，两种范式的性能都呈现下降趋势。特别值得注意的是，跨语言场景是所有测试中最具挑战性的，这说明**语言差异是来源判别系统面临的最大障碍**。

7 延伸研究：防检测与对抗

7.1 静态零样本干预

7.1.1 实验设计

基于对判别机制（Top 3 关键特征：段落数、行内公式数、段落长度）的深入理解，我们设计了一系列对抗性提示词，要求大模型在生成时主动规避” 暴露 AI 身份” 的结构特征。

7.1.2 实验结果

表 13: 防检测实验结果：各模型伪装成功率

模型	样本数	误判数	成功率
Deepseek	44	30	68.18%
Kimi	50	9	18.00%
GLM	50	35	70.00%
Qwen	50	49	**98.00%**
总计	**194**	**123**	**63.40%**

总体伪装成功率高达 63.40%。Qwen 在针对性提示词驱动下，伪装成功率达到惊人的 98%。

7.1.3 误判机制分析

表 14: 成功伪装 vs 被识别样本的特征对比

特征	成功伪装	被识别	差异
平均段落数	1.36	8.68	-84.3%
祈使代词密度	1.29	4.08	-68.4%
行内公式数	3.21	12.45	-74.2%
过渡词密度	0.85	2.67	-68.2%

数据清晰地表明，误判并非随机发生，而是因为这些样本确实被提示词" 推动" 到了更接近人类分布的区域。

7.2 动态迭代对抗实验（LLM-as-Optimizer）

7.2.1 实验架构

为了进一步探索 AI 自我进化的反侦察能力，我们设计了基于 LLM 反馈的迭代对抗实验：

- **判别器**：已训练好的分类模型（五分类器，准确率 95.5%）
- **生成器**：Kimi-k1.5（静态实验中最难被攻陷，伪装成功率仅 18%）
- **优化器**：DeepSeek-V3（负责将判别器反馈转化为 System Prompt 优化）
- **反馈循环**：生成器输出 → 判别器提取失败特征 → 优化器重写生成器的 System Prompt→ 下一轮生成

7.2.2 阶段一：先验主导型 Optimizer 的失败

在初始实验中，我们允许 DeepSeek" 自由发挥"，基于其对" 人类写作风格" 的先验理解来优化提示词。然而，这种策略遭遇了严重的" 幻觉叠加" 问题：

1. **刻板印象陷阱**：DeepSeek 认为" 人类就一定要写口水话"，于是在 Prompt 中盲目提出需要引入"umm"、"wait" 等非自然的口语化词汇。
2. **Prompt 质量崩塌**：生成的 System Prompt 逐渐演变成" 滑稽的舞台剧剧本"，充斥着相互矛盾的指令。
3. **性能崩溃**：Kimi 的绕过率止步于 33.33%，并在后续迭代中由于 Prompt 混乱而彻底崩溃。

7.2.3 阶段二：数据驱动型 Optimizer 的突破

基于阶段一的教训，我们对实验进行了改进，提示 DeepSeek 根据分类器的特征进行提示，强制其进行纯数据驱动的规则转化。于是 DeepSeek 给出了以下风格的 prompt：

> **Format**: Write exactly 1-2 paragraphs. Do not use lists, bullet points, or numbered steps.
> **Math Expressions**: Limit inline math expressions to a maximum of 5.
> **Banned Words**: Do not use "we", "let", "suppose", "because", "therefore", "firstly", "moreover"...

7.2.4 实验结果

表 15: 数据驱动型迭代对抗实验日志

轮次	核心策略	绕过率	状态
第 1 轮	基础提示词（无优化）	0/15 (0%)	基线
第 2 轮	结构化硬约束 Prompt	15/15 (100%)	**突破**

在接受了新的结构化提示词后，Kimi 生成的解答**瞬间实现了 100% 的检测绕过率**！

表 16: 先验主导 vs 数据驱动 Optimizer 对比

维度	先验主导型	数据驱动型
优化自由度	高（允许发散）	低（严格约束）
Prompt 质量	混乱、矛盾	清晰、可执行
迭代轮数达到 100%	失败（>8 轮）	**2 轮**
最终绕过率	33.33%	**100%**
关键问题	幻觉叠加	无

7.3 对抗判别器的实用建议

基于本研究的发现，我们总结出以下让 LLM 生成的数学作业更像” 人类学生手写” 的实用建议：

1. **控制段落数量**：将解答控制在 3-5 个段落，避免过度分段
2. **减少列表使用**：避免使用”Step 1/2/3” 或”Firstly/Secondly” 等机械列表
3. **降低公式密度**：适当减少行内公式的使用频率
4. **自然过渡**：使用” 注意到”、” 不难发现” 等自然语言过渡
5. **保留瑕疵**：适当保留一些人类常见的笔误或不规范表述

8 结论

本研究系统探讨了基于文本与逻辑特征的数学作业来源判别问题，完整回应了课程作业要求的三个核心问题：

问题一（特征提取）：我们提出了” 双轨特征工程” 框架，提取了 28 维结构特征与 TF-IDF 词汇特征。特征消融实验验证了该架构的协同效应，数学/LaTeX 特征被证明是最重要的子模块。

问题二（判别模型）：Hist Gradient Boosting 在同课程测试集上达到 95.5% 的五分类准确率，DistilBERT 达到 98.1% 的巅峰表现。

问题三（泛化性）：同分布域内泛化（95.5%）、跨学科跨域泛化（79.08%）、跨语言零样本泛化（54-57%）。跨学科实验发现 GLM 风格” 坍缩” 现象，跨语言测试揭示了语言差异是最大障碍。

延伸研究（防检测）：静态干预可使 LLM 伪装成功率达到 63.40%，数据驱动型反馈可使最难伪装的 Kimi 在 2 轮迭代内实现 100% 绕过率，揭示了当前检测方法在对抗场景下的根本性脆弱性。

参考文献

[1] Gehrmann, S., Strobelt, H., & Rush, A. M. GLTR: Statistical Detection and Visualization of Generated Text. In *ACL*, 2019: 111-116.

[2] Tian, E. GPTZero: An AI Text Detection Tool. `https://gptzero.me`, 2023.

[3] Sadasivan, V. S., et al. Can AI-Generated Text be Reliably Detected? *arXiv preprint arXiv:2303.11156*, 2023.

[4] Wang, H., et al. On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective. *arXiv preprint arXiv:2302.12095*, 2023.

[5] Stamatatos, E. A Survey of Modern Authorship Attribution Methods. *JASIST*, 2009, 60(3): 538-556.

[6] Zellers, R., et al. Defending Against Neural Fake News. In *NeurIPS*, 2019: 9054-9065.