一种用于多阶段理赔申请文档理解的混合架构：结合视觉语言模型与机器学习，实现实时处理

Lilu Cheng、Jingjun Lu、Yi Xuan Chan、Quoc Khai Nguyen,

John Bi、Sean Ho

人工智能团队，Fullerton Health

lilu.cheng@fullertonhealth.com

摘要

理赔文件是医疗保健和保险业务的基础，是报销、审计和合规工作的依据。然而，这些文件通常并非以数字格式生成；它们往往是扫描版 PDF 或在不受控条件下拍摄的照片。因此，文件内容存在显著差异——从打印发票到手写医疗报告，语言也各不相同。Fullerton Health 的业务充分体现了这一挑战：该公司每年在九个市场（包括新加坡、菲律宾、印度尼西亚、马来西亚、中国大陆、香港、越南、巴布亚新几内亚和柬埔寨）处理数千万份理赔申请。这种多样性，加上图像质量参差不齐、版面布局各异，给自动解析和结构化信息提取带来了重大挑战。

本文介绍了一种稳健的多阶段处理流程，整合了多语言光学字符识别（OCR）引擎 PaddleOCR、传统逻辑回归分类器以及紧凑型视觉语言模型（VLM）Qwen 2.5-VL-7B，以实现高效且

准确地从海量理赔数据中提取字段。所提出的系统实现了超过 95% 的文档类型分类准确率和约 87% 的字段级提取准确率，同时将平均处理延迟控制在每份文档 2 秒以内。相比之下，人工处理每项理赔申请通常需要约 10 分钟，我们的系统显著提升了效率。300× 这些结果表明，将传统机器学习模型与现代视觉语言模型相结合，能够在真实场景自动化应用中实现达到生产级水平的准确率和速度。该方案已成功部署于我们的移动应用程序，目前每周处理来自越南和新加坡的数万份理赔申请。

1 引言

保险理赔处理仍然是一项劳动密集且效率低下的工作。这源于工作流程的复杂性：人工操作员不仅要审核文档，还要进行信息提取、保单核验和最终裁定。因此，单个操作员的处理能力受到严重限制，形成了处理大量理赔申请的瓶颈。准确理解文本并提取信息是这一流程中的关键环节，但目前效率仍然不高。这个瓶颈影响的不仅是专业管理人员，也包括理赔申请人。例如，在通过移动端提交理赔申请时，用户必须手动填写医疗记录中的特定字段。然而，在小小的移动界面上查找并核实这些信息颇具挑战，尤其是存在认知差距：用户通常不熟悉出院小结等异构文档的结构和术语。因此，申请流程效率极低，每份理赔申请通常至少需要十分钟。事实上，人们已经尝试过许多基于规则的方法来解决

这一问题（Goolla，2025年；Kumar 和 Sharma，2024年）。然而，这些文档的异构性——不同的模板、语言和版面布局——阻碍了这些努力。近年来，随着大语言模型的发展，尤其是能够理解文档上下文的大型视觉语言模型（视觉语言模型）应运而生。这使得理解文档并自动从理赔文件中提取所需字段成为可能。尽管大型视觉语言模型具备准确提取所需的高级语义理解能力，但它们往往不适用于现实世界中的理赔处理。敏感医疗保健数据的处理对模型选择施加了严格限制，许多性能出色但存在隐私风险的外部模型因此无法使用。此外，大型视觉语言模型高昂的推理成本也阻碍了可扩展性。因此，挑战在于弥合性能差距：在紧凑、隐私保护框架的限制下，实现大型模型复杂的理解能力。

本文介绍了一种多阶段、资源高效的架构，该架构利用紧凑型视觉语言模型和多语言光学字符识别技术，从扫描理赔文件（图 1）中提取结构化字段。我们的设计将任务分解为预处理、混合分类、自适应提取和后处理，每个阶段都针对稳健性和可解释性进行了优化。图 1展示了所提出架构的概览。

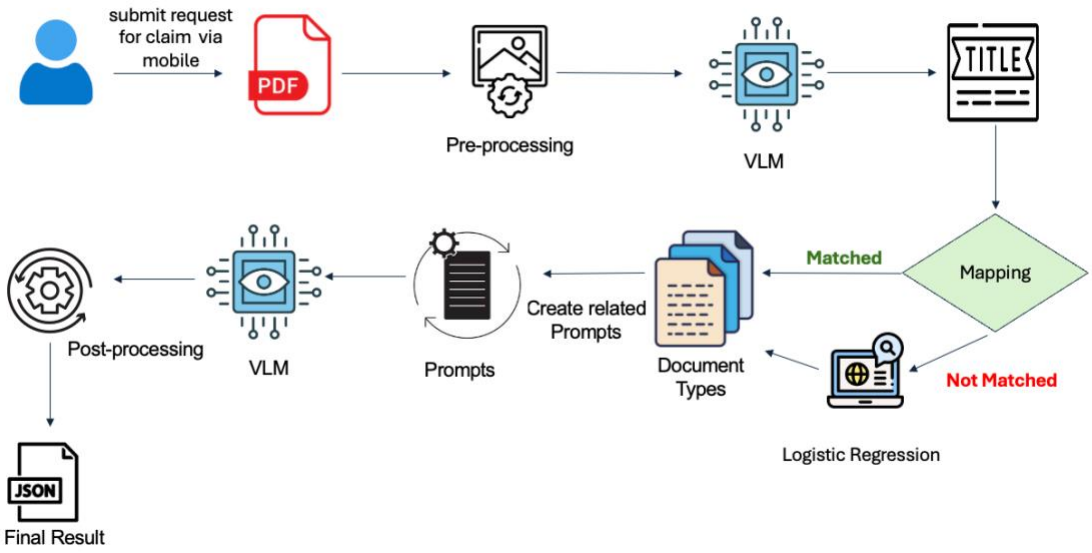

图 1 **多阶段理赔单据抽取流程的架构**。工作流程如下：(a) **预处理**：将原始提交材料分段并栅格化为图像。(b) **混合分类**：主视觉语言模型提取文档标题，以便进行基于规则的映射；对于无法映射的情况，则使用逻辑回归模型作为备用方案。(c) **自适应提取**：根据识别出的文档类型生成特定提示词，引导辅助视觉语言模型提取目标字段。(d) **后处理**：通过 Elasticsearch 对提取出的实体进行规范化，以实现知识库对齐。

2 背景

文档理解领域的研究通常分为两个核心方向：解析和信息抽取。

OCR 中的解析主要涉及解读文档布局并识别文本内容。早期一代 OCR系统（例如Casey 和 Lecolinet，1996；Mori 等人，1999）依赖人工设计的特征和基于规则的启发式方法。这些方法在受控条件下表现尚可，但难以应对现实世界文档的多样性和复杂性。深度学习的兴起—尤其是卷积神经网络（CNNs）及其后续变体—标志着 OCR 转向数据

驱动型 OCR（Goodfellow 等人，2014年；Shi 等人，2015年）。这些模型显著提升了识别准确率、稳健性和适应性。然而，随着人工智能应用不断扩展，OCR 技术开始面临新的需求：必须处理范围广泛得多的输入，包括手写文本、多语言文档、罕见文字系统，以及融合了表格、图表和嵌入图像的版面布局。在工业界和学术界，OCR 已从独立的识别任务发展为端到端智能系统中下游应用的关键赋能技术，例如文档理解、关键信息抽取（KIE）和语义检索。随着大语言模型（LLMs）和检索增强生成（RAG）系统的迅速发展，OCR 的作用变得愈发关键。这些系统依赖大量准确、多样且结构良好的文本数据进行训练和推理。在此背景下，OCR 不再只是文本数字化工具，而是推动整个知识处理流水线的基础组件—从将历史档案转化为可检索文本，到支持在多模态文档集合中进行实时问答。OCR 输出的精确性和完整性直接影响基于大语言模型的应用的可靠性和性能，尤其是在法律、学术或商业文档等以扫描件或基于图像的材料为主的领域。现代轻量级光学字符识别框架，例如 PaddleOCR v3（Cui 等人，2025年），现已支持多语言、旋转感知识别和层次化文档解析，即使面对低质量扫描件也能保持较高准确率。与此同时，布局感知型 Transformer 架构—包括 LayoutLM（Xu 等人，2020年）、LiLT（Yang 等人，2021年）和 DocLLM（Yao 等人，2024年）—通过联合建模文本和空间特征，更有效地捕捉文档的底层结构，进一步推动了该领域的发展。

相比之下，信息抽取旨在将非结构化文本转换为结构化数据。传统方法主要依赖基于解析结果构建的规则或关键词匹配方法。这些方法在处理整洁的单语言文档时效果显著，但难以应对

多语言内容和手写输入，而简单的关键词匹配方法则会失效。为克服这些局限，近期研究利用了能够理解整份文档语义的视觉语言模型（VLM）。GPT-4o（OpenAI，2024年）和 Qwen-2.5-VL（Bai等人，2025年）等模型在语义理解和多模态推理方面展现出卓越的能力。然而，在医疗保健、金融等领域，数据隐私限制往往禁止将敏感文档传输至外部视觉语言模型服务，即使 GPT-4o 或 GPT-o4 等大型专有模型具有更强的视觉能力。因此，MiniCPM-o 2.6、InternVL2.5-8B、Qwen2.5-VL-7B 等可在私有或本地部署环境中运行的开源视觉语言模型，提供了一种切实可行的替代方案。在这些模型中，Qwen-2.5-VL 7B 在性能与部署能力之间实现了良好平衡——它可部署在单张 NVIDIA A100 GPU 上，因此特别适用于本地化医疗应用。

3 方法

令 $\mathcal{T}$ 表示理赔文件类型的有限集合（例如理赔申请表、发票、收据、医疗报告）。每种类型 $t \in \mathcal{T}$ 对应一个模式 $\mathcal{S}_t = \{f_1(t), f_2(t), \cdots, f_m(t)\}$，该模式定义了信息抽取所需字段的集合，其中每个字段 $f_i(t)$ 都定义了要抽取的具体项目，例如患者姓名、保单号。

将文档实例记为

$$x = \{p_1, p_2, \cdots, p_n\},$$

其中每个页面 p_i 都代表一张扫描图像或 PDF文件中的渲染页面。页面可能包含印刷内容、手写内容或表格数据，也可能使用多种语言。任务是自动推断文档类型并提取特定于模式的字段。

文档类型分类

识别文档的语义类型：

$$t = g(x), \text{ where } t \in \mathcal{T} \ .$$

模式条件字段提取

对于已识别的文档类型 $\hat{t}$，提取 $\mathcal{S}\hat{t}$ 中定义的相应字段：

$$\hat{y} = h(x, \hat{t}) = \{(f, \hat{v}_f, \hat{b}_f, \hat{c}_f) | f \in \mathcal{S}_{\hat{t}}\}$$

其中，$\hat{v}_f$ 表示提取的字段值，$\hat{b}_f$ 表示证据（例如边界框或区域），$\hat{c}_f$ 表示提取置信度。

目标是在最大限度提高文档类型分类准确率和字段级提取质量的同时，尽量降低系统延迟，并确保可解释性和可审计性：

$$\max \left(Acc_{type}(g) + FLA(h) \right) \ \ s.t. \min Latency(g, h)$$

为满足这些要求，我们提出了一个四阶段处理流程：(1) 预处理，(2) 混合分类，(3) 自适应提取，(4) 后处理。

3.1 预-处理

对于理赔文件，大多数输入都是扫描件或照片。其中许多照片质量很高，因此文件体积较大（最高可达 50 MB），这会显著拖慢下游视觉语言模型的信息抽取

流程。因此，有必要设置一个预处理阶段，其中包括图像切分和调整尺寸。

在传统的 OCR 系统中，文档旋转时文本识别往往会失败（Goodfellow 等人，2014年；Shi 等人，2015年）。然而，由于 PaddleOCR v3 已具备图像质量增强及失真或方向调整系统（Cui 等人，2025年），即使输入图像发生旋转，它仍然稳健（我们的实验已验证这一点）。因此，在我们的处理流程中，不会处理方向问题，而只进行两个简单的预处理步骤：拆分和调整大小：

拆分 – 加载 PDF文件，并将每一页转换为图像。随后，这些页面图像将作为 PaddleOCR v3 和 Qwen-2.5-VL-7B 模型的输入。

调整大小 – Qwen-2.5-VL 7B 模型支持广泛的分辨率输入。默认情况下，它使用输入图像的原生分辨率，但更高的分辨率可以提升性能，代价是需要更多计算资源。用户可以设置最小和最大像素数，以获得适合自身需求的最佳配置，例如将令牌数量范围设为 256-1280，从而平衡速度和内存使用量。在我们的系统中，我们会将大图像缩小至 1024 像素的分辨率：

- 确保文档文本、表格和手写内容具备足够的视觉细节，同时不会产生过多开销，并且
- 符合模型在令牌预算权衡方面的最佳表现，从而在控制延迟和内存消耗的同时，实现良好的提取准确率。

预处理后，每个页面都会经过OCR系统，以恢复文本内容。

我们采用了 PaddleOCR v3框架，因其具备出色的多语言覆盖能力、较高的速度，以及处理含噪或扫描文档的强大能力。正如上文所述，PaddleOCR V3 由两个主要组件构成：

文本检测 – 可微边界框检测器可在多尺度下识别文本区域。这一能力对于复杂的财务文件或理赔文件尤为重要，因为其中包含叙述性文本、表格和边注等混合版式。

文本识别——随后使用特定语言识别头将检测到的区域转写为字符序列。

它支持 80 多种语言，包括英语、简体中文、越南语和繁体中文，适用于印刷体和手写体。对于每个识别标记，系统不仅会输出转录文本，还会输出置信度分数以及其边界框的空间坐标（图 2）。这些元数据使下游模块能够过滤掉低置信度结果，并保留表格和表单等版式敏感结构。在我们的解决方案中，我们需要利用置信度分数，引导下游用户更加关注低置信度结果。



图 2 使用 PaddleOCR 处理越南处方的示例

借助 PaddleOCR v3，我们的处理流程能够可靠地进行多语言文本转录并具备空间感知能力，为后续的信息抽取和分析任务奠定了坚实基础。最重要的是，我们可以将结果与视觉语言模型（VLM）的输出进行交叉核对，以提高整体准确率。

3.2 混合分类

理赔申请通常包含二十多种文件类型，例如理赔表格、处方、收据和发票，如表 1所示。此外，不同保险公司对所需文件有不同要求。通常，在移动理赔应用程序中，每个上传文件都被视为一种单独的文件类型。虽然有些理赔申请可能包含合并文件，其中包含多种文件类型，但这类情况将由一个高级系统处理，我们将在未来介绍该系统。因此，文档分类不仅对于将每个文件路由至适当的抽取模式至关重要，也有助于验证理赔完整性。一些保险公司以及 Fullerton Health 等第三方管理机构（TPA）要求每份理赔申请至少包含一份理赔表格，以及一张发票或收据。准确的分类可确保在理赔进入理赔裁定流程之前，所有必需文件均已提供。在本研究中，我们为文档中的每一页指定一种语义文档类型，例如理赔表格、发票、收据或医疗报告，以确定后续字段抽取所需的模式。为实现稳健的文档分类，我们将视觉语言模型表征与传统机器学习特征相结合，同时利用语义理解和基于版面的线索。

表 1 新加坡和越南的文件类型

	新加坡	越南
1	医院出院小结	医院出院小结
2	医院最终汇总账单	
3	医院最终明细账单	医院明细账单
4	医疗证明	医疗证明
5	发票	发票
6	收据	收据
7	转诊信	
8	担保函	担保函
9	X光报告	X光报告
10	诊断检查报告	诊断检查报告
11	实验室报告	实验室报告
12	处方	处方
13	组织学报告	组织学报告
14	CPF对账单	
15	理赔结算	理赔结算
16	入院前担保函 申请表	入院前担保函 申请表
17	检验申请单	检验申请单
18	医院入院前表格	医院入院前表格
19	理赔申请表	理赔申请表
20	初始担保函	初始担保函
21	最终担保函	最终担保函
22		出院证明
23		手术证明
24		出生证明
25		物理治疗记录
26		事故报告
27		车辆登记
28		驾驶执照
29		国民身份证
30		牙科治疗表

基于视觉语言模型的分类：首先使用 Qwen-2.5-VL-7B 模型处理每个页面图像，提示词为“识别文档类型：理赔表格、发票、医疗报告或收据。”并提供这些文档类型的定义。模型利用版面布局和视觉线索（例如标题、标志和结构化表单）来推断表单标题。在我们的实验中，直接使用视觉语言模型进行完整的文档类型

分类的准确率仅约为 70%，低于传统机器学习模型的 76%。主要限制在于所部署视觉语言模型版本的容量有限——文档类型分类所需的上下文推理能力超出了该模型的能力范围。因此，我们主要利用视觉语言模型抽取表单标题，因为标题通常包含强烈的语义线索，例如转诊信或出院小结。模型在这项抽取任务中的表现明显更好，因为标题区域通常是简短、界定清晰的文本片段，与视觉语言模型的自然语言理解和视觉定位能力十分契合。通过关注局部文本图像关系，而非对整页内容进行推理，模型能够更准确地解读文档页眉相关的字体大小、布局位置和语言模式。在大多数情况下，这种方法的准确率超过 95%；然而，发票或账单中的标题仍可能含义模糊、视觉上相似，因此更难区分。

机器学习–分类：为处理这些情况，我们使用 TF-IDF 嵌入作为输入，通过逻辑回归训练传统的机器学习分类器。将 2000 个样本按 80:20 的比例划分为训练集和测试集。在特征输入相同的条件下，我们比较了几种模型，包括随机森林、XGBoost 和逻辑回归。实证结果表明，逻辑回归在准确率和泛化能力之间取得了最佳平衡，因此被选为最终的文档类型分类器（表2）。

表2 传统机器学习模型的表现

机器学习模型	准确率（%）
XGBoost	76
随机森林	77
逻辑回归	87

3.通过紧凑型视觉语言模型实现自适应提取

分类后，系统使用 Qwen-2.5-VL-7B 触发针对特定模式的字段抽取。每种模式都定义了一组固定的目标字段。例如：

C理赔申请表：理赔编号、患者姓名、保单号、理赔金额 t

发票：服务提供方、日期、总金额，

医疗报告：诊断、服务提供方、医生姓名、入院日期

收据：收据编号、服务提供方、已付金额、付款日期

提示生成器会动态构建供模型使用的自然语言指令，例如：“从这张发票图像中提取以下字段：服务提供方名称、服务日期和总金额。以 JSON 格式返回结果。”

在实践中，我们发现，使用由以下四个部分组成的单槽结构化提示，能够显著提高提取准确率：

角色定义 – 指定模型的功能或视角（例如：“你是一名信息抽取助手……”）。

字段定义 – 清晰描述目标字段及其语义边界。

输出格式规范 – 定义预期的数据模式和输出样式（例如，JSON 或键值对）。

预期输出示例 – 提供参考示例，以指导模型的响应结构和语气。

紧凑型视觉语言模型联合处理视觉布局和 OCR词元，使其能够推断字段标签与相应字段值之间的语义关系。模型输出字段值以及

边界框和置信度分数，这些信息由 PaddleOCR 提供，以实现空间可追溯性。

与大型视觉语言模型相比，紧凑版本可部署在单台 GPU 服务器（如 H100）上，同时保持具有竞争力的准确率。

提取的信息以 JSON 格式导出，如下所示：

{ "claim_id": "C2024-0001", "
patient_name": "ABC", "policy_no":
"VN111", "diagnosis": " 急性支气管
*炎", "provider": "*河内综合医院", "
*visit_date": "2024-10-05", "total_*
amount": 1650000, }

3.4 后处理

尽管视觉语言模型能够生成高度准确的结果，但仍可能出现一些错误。例如，医院名称有时可能不一致或提取错误。为解决这一问题，我们采用基于 Elasticsearch 的规范化模块（Gormley 和 Tong，2015年），将提取出的每个医院名称替换为预定义参考列表中最相似的条目。为便于执行此流程，我们预先存储了一个医院名称数据库，收录来自卫生部、保险服务提供方网络等经核实来源，以及我们自己的

内部系统记录。在后处理过程中，系统使用模糊字符串匹配和相似度评分，将每个提取出的名称与此数据库进行比对。随后，系统会将置信度较低或存在歧义的结果替换为最接近的标准化医院名称。这确保了文档间的一致性，并显著提高了实体链接、理赔匹配和分析报告等下游任务的准确率。

此外，日期格式有时并不稳定。虽然预期格式为 DD/MM/YYYY，但偶尔会出现包含额外时间部分的输出（例如 DD/MM/YYYY HH:MM）。因此，应用日期格式验证规则，以确保所有输出的格式保持一致。

4 结果

我们使用从新加坡和越南收集的 4,200 份多语言理赔文件（约 32,000 页）评估了所提出的处理流程，涵盖 20 多种文件类型，例如理赔表格、发票、医疗报告和收据（表1）。数据集由英语（50%）和越南语（50%）文件组成，其中包含印刷内容（70%）和手写内容（30%）。评估指标包括文件类型分类准确率、字段级提取准确率和单份文件平均延迟。如表3所示，所提出的方法实现了 87% 的字段级准确率和 97% 的分类准确率。此外，单份文件的平均处理延迟为 2 秒，显示出计算效率的大幅提升。相比之下，人工处理每份文件通常需要约 10 分钟，这意味着我们的系统在效率方面提升了超过 300×。

表3 新加坡和越南理赔数据集的实验结果

	分类 准确率(%)	信息抽取 准确率(%)	平均处理时间 每份文件
SG	93	87	1.8
VN	97	75	2

尽管取得了这些出色的结果，仍存在若干误差来源。

首先，低对比度或手写文本会产生 OCR噪声，导致词元丢失或识别错误。例如，在一些使用点线的越南理赔表格中，手写内容——尤其是理赔金额——常常会被误读。最简单的解决方案是为客户提供空白表格填写，从而提高文本清晰度和 OCR准确率。

其次，日期歧义仍然是一项挑战。当同一份文件中出现多个日期（例如就诊日期和治疗日期）时，视觉语言模型偶尔会选错日期。此外，紧凑型模型在区分非常相似的术语时，稳定性仍然有限。

凭借达到的准确率，我们将人工智能解决方案集成到理赔提交应用程序中。用户上传理赔文件后，系统会自动确定文件类型，并提取理赔政策要求的关键字段，实时填入申请表。界面仍然易于使用，用户可以检查并更正先前分析中发现的任何残余错误。因此，系统每周为用户节省数十万分钟，大幅提升了理赔提交流程的整体效率。



图3 OCR API 调用摘要截图。左侧面板显示响应时间，大多数请求在两秒内完成。少数较长的响应时间可能是由于文档大小较大。右侧面板展示了 API 调用次数随时间的变化，概述了系统的使用情况和性能趋势。

5 结论与未来工作

我们提出了一种多阶段视觉–语言模型处理流程，用于实时处理中的全面理赔文档理解。该系统展现出高准确率、低延迟、强大的可扩展性和可解释性，并采用紧凑型模型，可高效部署在单台 GPU 服务器上，例如 H100 或 A100。该解决方案已作为 API 完全集成到我们的生产理赔系统中，每天为数万名用户提供服务，并在数秒内给出准确结果（图 3）。在实际应用中，我们的人工智能解决方案每周可节省数十万分钟，将每份文档的人工处理时间从 10 分钟缩短至约两秒。

展望未来，我们计划扩展系统的多语言能力，支持简体中文、繁体中文、马来语和印度尼西亚语等更多语言，并支持来自更多国家的理赔文件。 最终，我们的目标是扩大这一解决方案的应用范围，为数百万用户提供服务。文档理解是实现理赔处理全自动化的基础步骤，因此我们计划构建一个多智能体理赔裁定框架，利用所提出的多阶段字段提取流水线的输出。

致谢

这项工作由 Fullerton Health AI 团队与新加坡和越南地区的理赔运营、产品团队及工程团队合作完成。我们衷心感谢 Morrison Loh 博士、Ho Kuen Loon 先生和 Brian Lim 先生在整个工作过程中给予的宝贵指导、鼓励与支持。

参考文献

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J.,