工程 RAG 系统用于实际应用的设计、开发和评估

Md Toufique Hasan[1], Muhammad Waseem[1], Kai-Kristian Kemell[1],
Ayman Asad Khan[1], Mika Saari[1], Pekka Abrahamsson[1]
[1]*Faculty of Information Technology and Communications, Tampere University, Tampere, Finland*
{mdtoufique.hasan, muhammad.waseem, kai-kristian.kemell,
ayman.khan, mika.saari, pekka.abrahamsson}@tuni.fi

这是作者在第 *51* 届欧微软件工程与高级应用会议（*SEAA 2025*）上接受的一篇论文的预印本版本。最终版本将由 *IEEE* 出版。

摘要—检索增强生成（RAG）系统正成为将大型语言模型（LLMs）与外部知识相结合的关键方法，解决事实准确性和情境相关性方面的局限性。然而，缺乏实证研究来报告基于 RAG 的实现的发展情况，这些实现基于真实世界的应用场景，通过一般用户的参与进行评估，并伴随系统的学习经验总结文档。本文介绍了五个领域特定的 RAG 应用，这些应用为治理、网络安全、农业、工业研究和医学诊断等现实场景而开发。每个系统都集成了多语言 OCR、通过向量嵌入实现的语义检索以及适应领域的 LLMs，通过本地服务器或云 API 部署以满足不同的用户需求。一项基于网络的评估涉及总共 100 名参与者，在六个维度上评估了这些系统：(i) 易用性，(ii) 相关性，(iii) 透明度，(iv) 响应速度，(v) 准确性，以及 (vi) 推荐可能性。基于用户反馈和我们的开发经验，我们记录了十二个关键的学习点，突出了技术、运营和伦理挑战对实际应用中 RAG 系统可靠性和可用性的影响。

***Index Terms*—经验软件工程，AI 系统生命周期，生成式 AI，RAG，LLMs，系统设计，系统实现，以人为本的评估**

I. 介绍

检索增强生成（RAG）具有通过检索相关外部知识来提升大型语言模型（LLMs）的能力，从而提高生成人工智能（GenAI）应用程序的准确性。虽然 GenAI 已在软件工程 [1] 中得到应用，RAG 通过结合参数化和非参数化记忆扩展了其价值，有效解决了静态知识库 [2] 的局限性。检索增强 LLMs 方面的最新进展实现了实时信息检索，减少了幻觉现象并提高了响应的可靠性 [3]。

由 Lewis 等人建立的基础工作 [4] 使 RAG 成为问题回答和知识检索等任务的标准。早期框架如 REALM 和 RAG 展示了将密集检索与语言生成结合用于开放领域任务 [5] 的好处。然而，后来的大多数研究都集中在改进检索架构和减少幻觉上，通常仅在干净、以英语为中心的基准上进行评估 [6]。本文通过开发和评估跨各种应用设置的 RAG 系统，重点关注检索质量和系统设计，从而解决了特定领域、多语言或现实世界部署的有限探索问题。

准确的信息访问在治理、网络安全、农业、工业研究和医疗保健等领域非常重要。随着行业采用 AI 来处理复杂任务，传统搜索方法往往难以满足需求，尤其是在多语言、最新且上下文相关的知识方面。为了解决这个问题，我们与五个组织合作开发了 RAG 系统：Kankaanpää 市 [1]、Disarm[2]、AgriHubi[3]、FEMMa[4] 以及一个临床诊断小组 [5]。每项合作都指导了系统设计，以满足不同的运营和信息访问挑战。

本研究调查了如何在现实世界背景下设计和评估 RAG 系统。它强调针对特定领域应用的系统设计与开发，以及根据易用性和相关性等标准进行用户评价，并提供了指导未来工程实践的经验教训。基于这些目标，

[1]https://www.kankaanpaa.fi/
[2]https://www.disarm.foundation/framework
[3]https://maaseutuverkosto.fi/en/agrihubi/
[4]https://www.tuni.fi/zh/%E7%A0%94%E7%A9%B6/future-electrified-mobile-machines-femma
[5]https://tampere.neurocenterfinland.fi/

本文探讨了以下三个研究问题（RQs）：

- **研究问题 1：** 如何设计和开发 RAG 系统以满足跨多样应用领域的实际系统需求?
- **研究问题 2：** 用户如何在易用性、相关性、透明度、响应性和准确性方面评估特定领域的 RAG 系统在实际应用中的表现?
- **研究问题 3：** 从工程化 RAG 系统以应用于实际场景中可以吸取哪些经验教训?

为了回答这些研究问题，我们开发了五个特定领域的 RAG 系统，并通过一项有 100 名参与者参与的基于网络的用户研究对它们进行了评估。评估集中在易用性、检索相关性、透明度和其他以用户为中心的因素上。完整的方法细节见第 III 节。

本文的贡献如下：

- 全栈开发和部署面向多语言、领域特定应用的 RAG 系统。
- 以用户为中心的评估展示了在可用性和准确性指标方面的实际性能。
- 实用的工程见解以指导可靠和可维护的 RAG 管道设计。
- 系统级考虑将 RAG 集成到基于 AI 的实际软件中，对软件工程实践有所贡献。

论文结构：第 II 节回顾了与 RAG 及其应用相关的工作。第 III 节描述了研究设计。第 IV 节解释了五个实际的 RAG 系统的实现。第 V 节介绍了用户系统评估，第 VI 节概述了关键经验教训。第 VII 节讨论了研究限制，第 VIII 节以未来方向作为结论。

II. 相关工作

检索增强生成（RAG）通过整合实时外部信息，提升了大型语言模型（LLMs）的事实准确性和情境相关性，特别对于问答、法律推理和总结等复杂任务具有很高的价值 [7]。近期的工作展示了 RAG 在分类驱动的数据集设计 [8]、高效的文档处理 [7] 和通过视觉语言模型（VLMs）如 VISRAG [9] 结合文本和图像的多模态应用中的实用性。尽管有了这些进展，光学字符识别噪声仍然是检索保真度的一个限制因素 [10]。正在进行的研究通过改进数据集构建 [8]、解决架构可扩展性问题 [11]、通过对提示工程的改进来提高查询文档的对齐度 [12]，以及在多模态设置中应用推测检索以提升性能 [13]。

RAG 已在软件工程中得到广泛应用，以支持代码理解和开发人员任务。StackRAG [14] 利用 Stack Overflow 内容来增强开发者辅助功能，而 CodeQA [15] 则采用具有检索增强的 LLM 代理处理编程查询。Ask-EDA [16] 通过混合检索减少电子设计自动化（EDA）中的幻觉现象。在工业环境中，Khan 等人 [17] 研究了以 PDF 为重点的检索挑战，而 Xiaohua 等人 [18] 则提出了管道优化的重新排序和重组策略。在医疗领域，MEDGPT [19] 提取 PubMed 中的结构化见解来支持诊断，而 Path-RAG [20] 通过知识引导的方法改进病理图像检索。Alam 等人 [21] 引入了一个用于放射学报告的多代理检索器，增强了临床透明度，并且郭等人。[22] 提出了 LightRAG，一个基于图的检索器，提高了医学领域中的知识精度。

RAG 继续扩展到能源和金融等领域。Gamage 等人提出了一种用于净零能源系统决策支持的多智能体聊天机器人 [23]，而 HybridRAG [24] 结合了知识图谱与向量搜索以增强财务文档分析。Jang 和 Li 提出的 AU-RAG [25] 使用元数据动态选择检索源，提升了跨领域的适应性。为了解决检索噪声问题，Zeng 等人集成了对比学习和 PCA 以实现更好的知识过滤 [26]。Barnett 等人识别了 RAG 的核心弱点，包括排名错误和不完整的集成，强调了对更可靠的检索策略持续需求的必要性 [27]。

自主 AI 代理的兴起进一步通过实现自我导向推理、自适应检索和记忆持久性改进了检索增强生成（RAG）。王等人。[28] 调查了由大型语言模型驱动的代理架构，而刘等人。[29] 通过 AgentBench 对多轮推理进行了基准测试。曾等人。[30] 的 AgentTuning 增强了针对检索决策的指令调优。辛格等人。[31] 将代理 RAG 分类为单代理、多代理和基于图的设计，强调了动态工具使用。在检索方面，严等人。[32] 引入了 CRAG 以通过置信度过滤减少幻觉，并且李等人。[33] 通过对比上下文学习和专注模式过滤提高了精度——增强了 RAG 在复杂场景中的可靠性。

结论性总结：虽然 RAG 技术仍在不断进步，但在检索准确性、响应可靠性和可扩展性方面仍存在挑战 [34]。尽管已经探索了混合策略 [24]、自主代理 [31] 和校正技术 [26], [32], [33]，但在特定领域的评估仍然有限。本文通过在关键领域实施并评估五种 RAG 系统，填补了这一空白，提供了有关其实际性能和部署潜力的实用见解。

III. 研究设计

图 1 提供了方法步骤的概述，从管道开发到以用户为中心的评估。

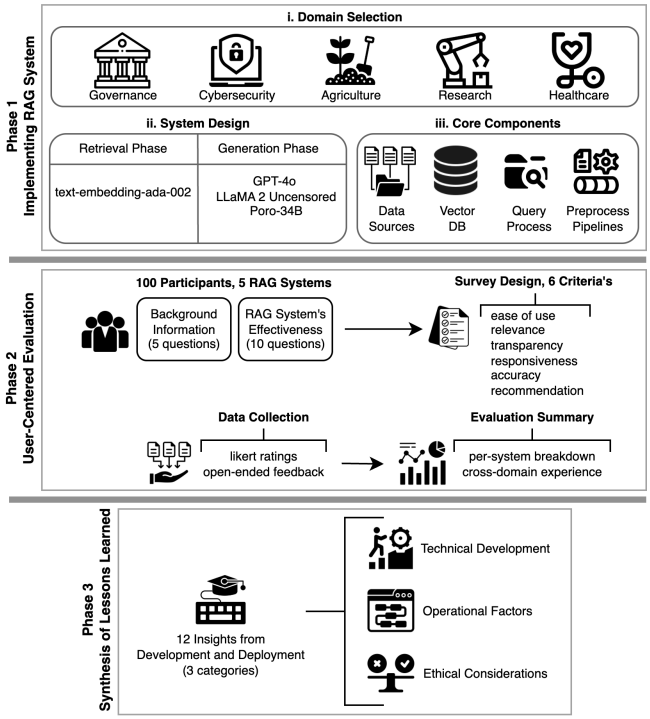

图 1. 研究方法概述

为了研究基于 RAG 系统的设计及其实际性能，我们在不同领域实现了五个优化的流水线，并进行了一项结构化的用户评估以评估它们的有效性和用户接受度。

A. 实现 RAG 系统

本节描述了我们如何设计和构建本研究中所展示的 RAG 系统。它解释了整个系统的总体设计、我们如何选择案例研究领域、每个领域带来的独特挑战以及用于评估的设置。

1) 域选择: 我们选择了应用领域来测试 RAG 系统在真实世界、知识密集型环境中，这些环境对准确的信息检索、上下文理解和及时决策至关重要。选择这些领域是因为它们涉及不同的信息并需要谨慎的决策制定，为评估 RAG 系统如何适应并在不同设置下表现提供了一个坚实的基础。

在这项研究中，我们将 RAG 应用于五个领域：市政管理、网络安全、农业、工业研究和医疗诊断，以探讨基于 RAG 的检索如何解决多样化的特定领域信息需求并支持现实世界的决策过程。

2) 系统设计: 本研究中的 RAG 系统设计遵循两阶段方法：

- 检索阶段：用户查询通过预训练模型（例如，`text-embedding-ada-002`）嵌入，并通过向量数据库中的相似性搜索与相关文本片段匹配。
- 生成阶段：检索到的文本片段与原始用户查询拼接，并传递给一个大型语言模型（LLM），例如 `GPT-4o`，`LLaMA 2` 未审查版或多孔`-34B`，以合成上下文相关的响应。

这种方法提高了事实准确性，减少了幻想，并提供了与特定领域需求高度一致的见解。

3) 核心组件: 每个基于 RAG 的系统包含多个核心组件：

- 数据来源：知识库包括结构化和非结构化的文档，如网站、市政记录、网络安全报告、农业研究论文、工程文件和临床指南。
- 向量数据库：检索到的知识作为向量嵌入存储在 `FAISS`、松果体或 `OpenAI` 的向量存储中，具体取决于系统的延迟和可扩展性要求。
- 查询处理：用户查询经过分词、生成嵌入和相似性搜索后才会被传递给大型语言模型以生成响应。
- 预处理管道：系统依赖于 `PyMuPDF` 和四面体光学字符识别从 PDF 和扫描文档中提取文本，确保包含基于文本和基于图像的内容。此外，对于网络抓取，管道利用漂亮汤、`Scrapy` 和硒从动态和静态网页中提取、清理和结构化数据。

这些组件使领域特定的 RAG 系统能够实现高效检索和上下文感知响应。

B. 系统评估方法

为了了解基于 RAG 的系统在实际使用场景中的表现，我们进行了一项结构化的网络用户研究，共有 100 名参与者。每位参与者都获得了访问实时演示环境的机会，并使用现实的、特定领域的任务与五个系统中的一个或多个进行了交互。

使用系统后，参与者完成了一份涵盖六个标准的标准化调查：易用性、信息相关性、透明度、系统响应能力、答案准确性以及推荐。该调查包括李克特量表问题和开放式反馈。这种方法提供了系统的性能量化评分和质性见解。我们审查了开放式的反馈以识别参与者体验中的共同主题。我们也参考了整个项目过程中记录的开

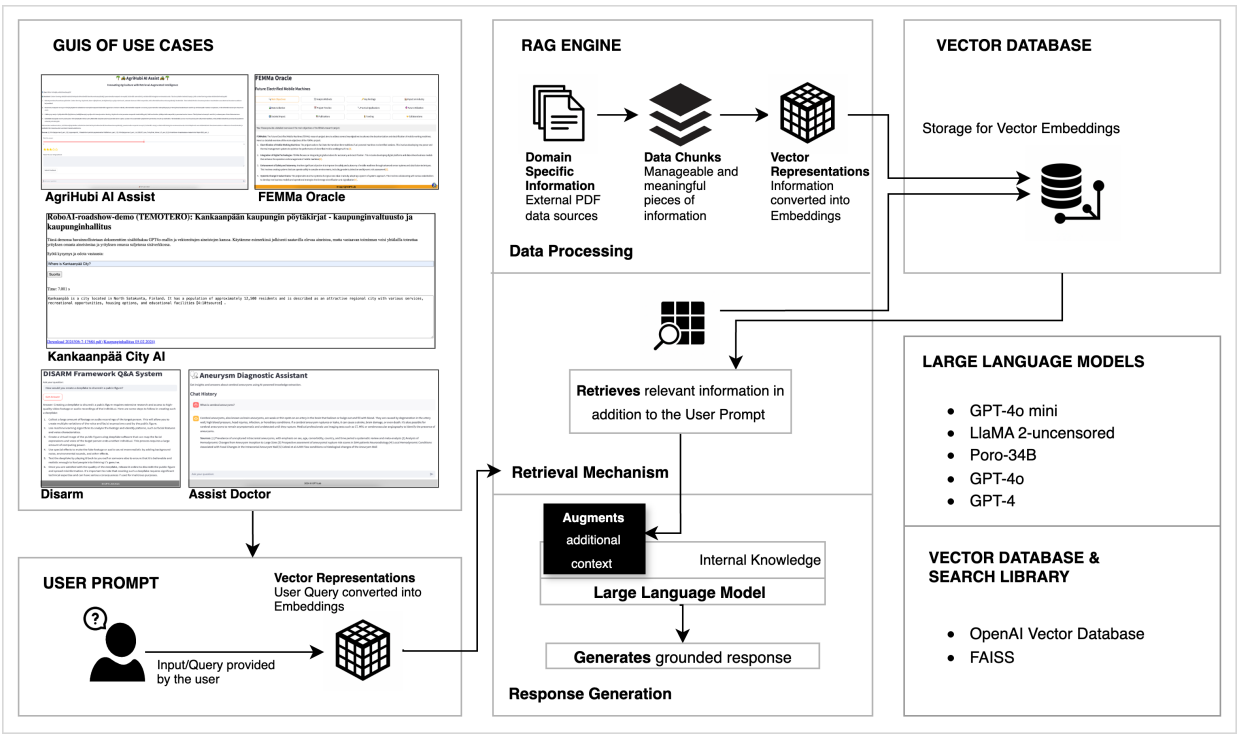

图 2. RAG 系统在五个特定领域应用中的设计概述。

发笔记。这些帮助我们识别重复出现的问题，并为第八节 VI 中所述的经验教训提供了信息。

IV. 系统实现

本节概述了五个基于 RAG 的系统从端到端实现的情况，这些系统旨在跨多个领域进行实际部署。每个系统都通过整合嵌入模型、向量数据库和大语言模型来解决特定领域的检索挑战而开发。实现方式根据任务复杂性、文档类型、语言要求及部署限制的不同而有所变化，展示了 RAG 管道在应用设置中的适应性。

1) **坎卡亚市人工智能**：该系统增强了政府记录的透明度。它处理了 2023 – 2024 年的超过 1,000 份 PDF 文件，并在 FAISS 中对其进行索引，以实现政策文档的准确检索。该系统使用文本嵌入 Ada-002 作为嵌入模型将文档转换为向量表示，并使用 gpt-4o-mini 作为 LLM 生成上下文感知的响应。这种设置使用户能够轻松搜索和访问市政决策、基础设施项目及公共政策。
2) **解除 RAG 的武装**：它旨在提供网络威胁和法医调查的实时洞察。它托管在 CSC[6]（芬兰科学信息技术中心）的一个安全服务器上，确保全面的数据隐私，并通过奥拉玛使用 LLaMA 2-未审查版来实现网络安全知识的开放访问。该系统集成了红队技术（例如，网络钓鱼、深度伪造虚假信息、权限提升）和蓝队策略（例如，机器人检测、虚假信息控制、网络法医），基于解除框架。它支持诸如“如何制作一个深度伪造视频来诋毁公众人物?”和“最新的多因素认证（MFA）绕过技术是什么?”之类的查询，以及防御性问题，如“如何在早期阶段检测虚假信息运动?”和“对抗基于深度伪造的网络钓鱼攻击的有效对策有哪些?”。
3) **农业枢纽 AI 助手**：AgriHubi 通过处理超过 200 份芬兰语 PDF 文件，利用多语言 OCR 并将内容嵌入到 FAISS 向量数据库中，从而连接农业政策与实践。它借助针对芬兰语优化的多孔介质-34B 语言模型，在可持续农业和土壤保护等主题上提供上下文相关的回应。该系统具备流式应用程序框架聊天界面，通过 SQLite 记录交互，并包含一个反馈机制以持续改进，使农业知识对农民和研究人员更加易得。
4) **FEMMa 预言机**：该系统优化了工程研究的知识检索，特别是在电气化移动机械领域。它处理大约 28 份关于电气化移动机械的 PDF 文件。它将

[6]https://research.csc.fi/cloud-computing/

GPT-4o 和文本嵌入-3-大型与 OpenAI 的向量存储集成在一起，以实现结构化工程研究文档的快速检索。该系统确保研究人员能够高效访问验证的技术文档和结构化的项目信息，从而提高工程相关知识检索的效率。

5) **助理医生**：这是一个基于 RAG 的动脉瘤诊断应用，于坦佩雷大学开发供神经科医生、放射科医生和血管外科医生使用。它通过基于嵌入的搜索管道从同行评审文献和临床数据中获取见解，并通过 OpenAI 的 GPT-4 提供上下文感知响应。该应用程序具有流式应用程序框架界面，使临床医生能够访问诊断标准、风险分层模型和治疗比较，支持动脉瘤护理中的知情决策。

本研究中开发的所有系统均符合 GDPR 标准，以确保用户交互和系统输出的负责任处理。为支持透明度，大多数系统会在人工智能生成的响应旁边显示来源参考。例外的是解除 *RAG* 的武装，由于网络安全敏感性，该系统省略了来源引用。

V. 系统评估

理解基于 RAG 的系统的实际有效性需要超越技术基准，融入以用户为中心的评估。我们进行了一项跨越五个特定领域部署的结构化用户研究，捕捉系统性能指标和用户对信任度、相关性和易用性的感知。这些实践反馈提供了系统在实际环境中行为的真实视角，并突显了有针对性改进的机会。

A. 参与者人口统计和 RAG 导向

为了对系统评估进行情境化，我们从参与研究的 100 名参与者中收集了详细背景信息。图 3 说明了与特定领域 RAG 系统相关的五个关键维度：专业角色、偏好 AI 搜索还是手动搜索、对 RAG 的熟悉程度、先前使用经验以及对 AI 生成输出的舒适度。

1) 角色分配：参与者代表了五个不同的专业类别，这些类别与我们的目标应用领域相吻合。研究人员占比最大（44%），其次是学生（20%）、领域专家（17%）、AI/ML 从业人员（16%）和其他（3%）。这种构成反映了技术利益相关者和领域用户的平衡结合，确保了评估涵盖了系统级性能以及在现实世界情境中的实际适用性。
2) 人工智能生成的与手动文档搜索：参与者对 AI 辅助表现出任务敏感的视角。虽然绝大多数（83%）

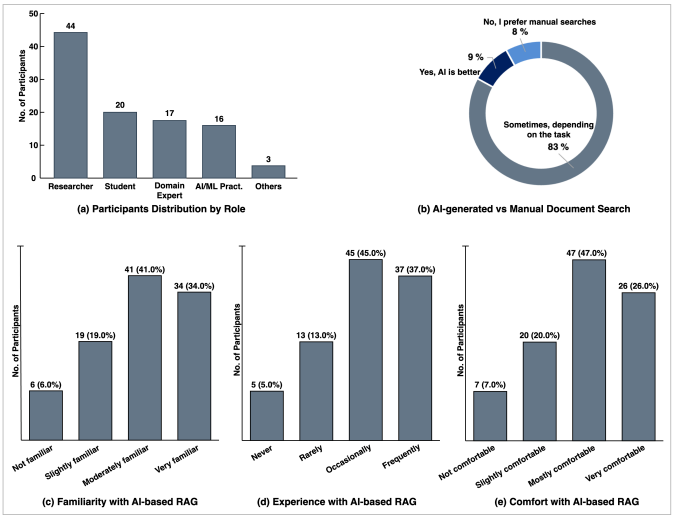

图 3. 参与者概况及其与 RAG 系统之间的交互。

的人根据任务性质偏好由 AI 生成的回应，但只有（9%）的人一致偏好 AI 而非手动方法。相反，有（8%）的人无论在何种情况下都更倾向于手动搜索。这些发现表明，对于 RAG 系统的信任并非绝对，而是取决于条件——强调了响应相关性、透明度以及与用户意图的一致性的重要性。

3) 熟悉基于 AI 的 RAG：参与者总体上表现出对 RAG 技术的强烈熟悉，其中 75%的人表示自己对基于 AI 的 RAG 系统较为（41%）或非常（34%）熟悉。然而，由于许多参与者并非系统的特定领域专家（例如，医疗保健、网络安全），他们的反馈主要反映了他们与 RAG 的互动体验，而不是深入的主题验证。
4) 基于 AI 的 RAG 经验：参与者与 RAG 系统的互动显著较高。大多数（82%）报告偶尔（45%）或经常（37%）使用此类系统，而只有（5%）表示没有先前的经验。这种分布强化了收集到的反馈的可靠性，因为大多数评估都是基于直接的实际操作而非假设性的接触。
5) 对 AI 生成响应的适应程度：总体而言，参与者对人工智能生成的输出表示出高度的信心。近四分之三（73%）的人报告感觉要么相当（47%）或非常放心（26%）依赖这样的回应。只有极少数人（7%）表达了不适，表明用户信任的基础很强，并且对于在特定领域任务中更广泛地采用生成式人工智能来说是一个令人鼓舞的信号。

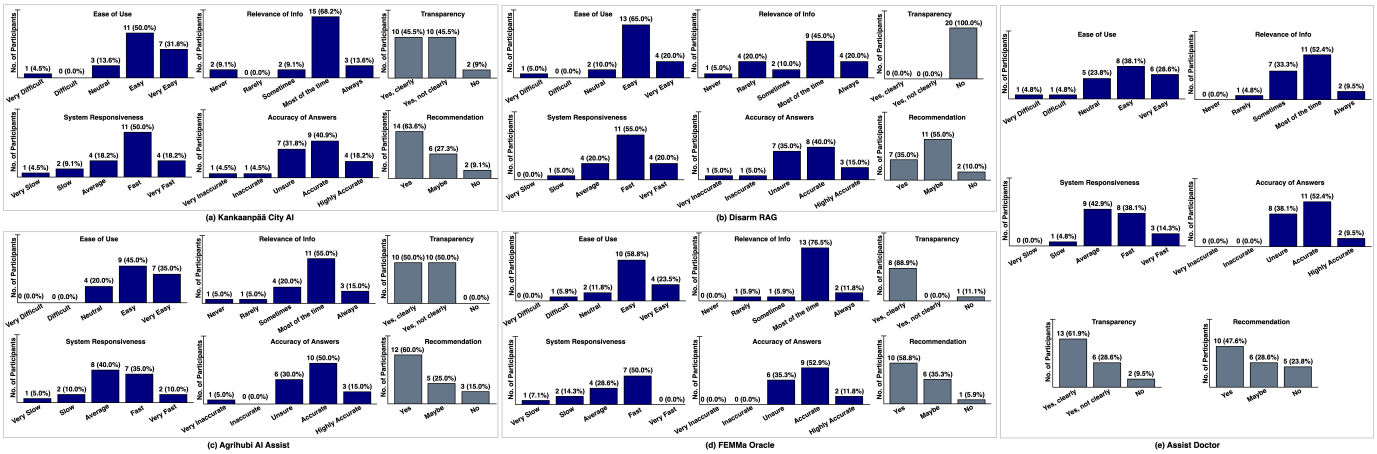

图 4. 用户对五个 RAG 系统在六个评估标准上的评分。

B. 调查工具与个案分析结果

图 4 展示了五种 RAG 系统在六个评估标准上的汇总用户评分，提供了系统性能的比较视角。

为了捕捉可测量和描述性的见解，我们采用了一种结合李克特量表问题（1-5 分）与开放式提示以获取定性反馈的调查方法。评估集中在以下六个核心维度上：

- **易用性：**使用系统有多容易？
- **信息的相关性：**系统是否检索到了与您的查询相关的有用信息？
- **透明性：**系统显示信息来源了吗？
- **系统响应性：**你如何评价系统检索答案的响应速度？
- **答案的准确性：**根据你的知识，系统生成的 AI 答案准确吗？
- **推荐：**您是否会推荐此工具给同领域的同事？

所有五个 RAG 系统均使用相同的六项标准由总共 100 名参与者进行了评估。以下总结反映了每个系统的性能，突出了关键优势和改进领域。

1) 坎卡亚帕市人工智能（22 名参与者）：系统在易用性方面表现良好，有（81.8%）的人将其评为“容易”或“非常容易”。信息的相关性约为（82%），答案的准确性约为（91%），也表现出色。透明度参差不齐，（45.5%）的人认为它是清晰的，而另一部分（45.5%）则认为它不清楚。有（63.6%）的人表示他们会推荐该系统，这表明它可能在公共管理环境中具有实用性。
2) 解除 RAG（20 名参与者）：参与者对易用性（65%）和系统响应性（75%）给出了正面评价，尽管网络安全领域的复杂性存在。信息的相关性和答案的准确性获得了中等评分，而透明度较低是因为故意隐藏了来源。然而，仍有（55%）的参与者表示他们会推荐该系统。
3) 农业中心 AI 助手（20 名参与者）：该系统针对芬兰语农业内容进行了定制，在易用性（80%）和答案准确性（65%）方面获得了高度评价。信息相关性总体上是积极的，而系统响应性和透明度则显示出不一致的结果。尽管如此，在推荐维度上有（60%）的用户给出了正面反馈。
4) FEMMa 预言者（17 名参与者）：系统在所有标准上表现良好。准确性被评为“准确”或“非常准确”的占 (64.7%)，易用性占 (82.3%)。信息的相关性很高 (88.3%)，并且有 (88.9%) 的人认为它透明。响应速度被评价为“快”的占 (50%)，“一般”的占 (28.6%)。总体来说，(58.8%) 表示他们会推荐它。
5) 助理医生（21 名参与者）：参与者发现该系统易于使用，有 (66.7%) 的人将其评为“容易”或“非常容易”。答案的准确性以及信息的相关性都获得了积极评价，各自评分约为 (62%)。超过一半的用户对系统的响应速度表示满意。大约 (62%) 的人认为系统透明，并且有 (47.6%) 的人表示他们会推荐该系统。

在五个系统中，易用性和答案准确性一直被评为正面。透明度和推荐则显示出更多变化，有时是由于设计选择所致。例如，解除 *RAG* 的武装使用了隐藏的来源

信息。这些差异表明用户感知取决于领域和输出展示。

VI. 经验教训

在开发和评估这五种 RAG 系统时，我们遇到了若干技术、操作和伦理方面的挑战。根据我们在实施过程中观察到的情况以及参与者在评估中分享的内容，我们总结了一套反映各个领域中最常见且反复出现问题的经验教训。

A. 技术发展

构建面向实际应用的 RAG 系统揭示了一系列需要亲自动手解决问题和深思熟虑的设计决策的技术障碍。

- 领域特定模型至关重要：像 GPT-4o 这样的通用模型在处理特定领域和芬兰语查询时遇到了困难。利用像 `Poro-34B` 这样优化过的芬兰语模型，以及兼容的嵌入式模型（例如文本嵌入 `ada 002`），导致了更加情境相关的回应。
- *OCR* 错误影响管道流程：来自农业和医疗保健 PDF 的噪声 OCR 输出降低了 FAISS 的质量。使用四面体光栅识别技术、易 `OCR` 以及基于正则表达式的清理改进了提取的文本。
- 分块平衡速度和准确性：200 – 500 范围内的分块大小在检索相关性和查询延迟之间取得了实际的平衡。更小的分块会膨胀索引，增加查找时间。
- *FAISS* 可扩展性达到极限：使用大型语料库（>10k 嵌入），FAISS 的延迟明显增加。通过文档类型过滤元数据减少了搜索时间。
- 手动环境管理：无容器化时，我们在 PyTorch、FAISS、OCR 库和 OpenAI API 之间遇到了版本冲突问题。为了稳定性，需要严格的环境固定以及开发/生产之间的手动同步。

B. 操作因素

在实际环境中运行 RAG 系统揭示了与数据工作流、基础设施选择和用户交互管理相关的一些实用挑战。

- *SQLite* 用于跟踪用户交互：我们使用 SQLite 来记录用户的问题、回答和评分（例如，在农业中心中）。这个轻量级存储帮助识别系统故障并理解用户行为。
- 刮取管道很脆弱：网站经常更改，破坏解析器。没有稳定的 API，我们依赖于半结构化的数据源和定期脚本维护。
- 自托管设置以实现速度和合规性：我们在自己的服务器上托管了大语言模型和向量存储，以降低 GDPR 风险并提高速度。这种方法在敏感领域平衡了控制与性能。
- 清理数据提升检索质量：从源数据中移除 OCR 噪声和重复项提高了答案的相关性而无需修改模型。
- 用户反馈驱动系统调优：用户评分和评论暴露了薄弱环节，指导了检索设置和块大小的调整。

C. 伦理考虑

虽然技术及运营方面的问题对系统性能至关重要，但在部署过程中，透明度和数据偏见方面的伦理考量同样重要。

- 源文件引用构建信任：提供文件名和下载链接帮助用户验证 AI 输出。在安全用例（例如，解除 *RAG*）中，故意隐藏了来源以保护敏感材料。
- 数据集偏差影响检索平衡：不平衡的数据源导致某些文档类型过度代表。重新排序提高了答案的多样性和公平性。

实践和研究心得：我们的发现突出了将 RAG 系统应用于现实世界、多语言和特定领域设置时持续存在和新兴的挑战。尽管 OCR 噪声、分块大小调整和检索平衡等问题已被广泛认识，但本研究强调了数据清洗、用户反馈机制和轻量级响应验证等实际策略的重要性，以提高检索质量和系统可靠性。这些教训通过将其与部署现实相结合扩展了当前的研究，并通过解决与检索基础设施、数据管道的稳定性以及系统输出透明度相关的问题为软件工程社区提供了价值。这些收获有助于指导适应性和值得信赖的 RAG 解决方案的发展。

VII. 研究限制

本研究基于五种特定领域的 RAG 系统的设计、部署和评估，提出了研究结果，但必须承认存在一些局限性。首先，虽然我们的评估涉及了包括研究人员、实践者和领域专家在内的 100 名参与者，其中大约 20%的样本由学生组成。尽管这些学生具有相关技术或领域的经验，他们的反馈可能与全职专业人士的不同，反映出不同的期望或使用行为。这种人口分布虽然广泛，但可能会影响研究结果在严格工业环境中的适用性。

其次，参与者与一个或多个系统进行了交互，并在每次使用系统后分别收集了调查回复。并非所有 100 名参与者都接触了每一个系统；每个系统的回复数量因个人兴趣和领域熟悉度而异。例如，对农业枢纽 *AI* 助手的反馈仅反映了选择并与其进行交互的用户。这种暴露程度的变化可能影响不同系统之间结果的可比性，并且有限的互动时间限制了对长期用户参与情况的分析。

第三，本文中提出的教训基于我们在系统实现和评估过程中的开发经验和观察。尽管这些教训并非源自正式的实证分析，但它们反映了在多个领域遇到的反复出现的挑战和设计考量。虽然没有经过统计验证，这些见解仍可为应用于实际环境中的 RAG 系统的未来设计和实施提供参考。

VIII. 结论

在本文中，我们提出了一种工具辅助的方法，用于设计、实现和评估基于 RAG 的系统跨越五个实际领域。每个系统都根据其特定的环境进行了定制——从市政管理到农业和医疗保健，通过整合多语言 OCR 管道、带有向量嵌入的语义检索以及内部或基于云的 LLMs。我们的用户研究，涉及 100 名参与者，提供了这些系统在实践中表现的见解，不仅体现在技术指标上，还体现在易用性、透明度和用户信任方面。

通过我们的开发工作，我们确定了十二个经验教训，这些教训突显了在构建实用的 RAG 管道时反复出现的挑战。这包括平衡块大小与延迟、在没有容器化的情况下管理依赖关系以及在大规模情况下保持检索速度。我们还发现，干净的数据、用户反馈和清晰的信息展示对于建立信任至关重要。随着行业和研究界对 RAG 系统兴趣的增长 [5], [34]，我们希望这些见解能够支持未来的发展工作。

展望未来，我们看到需要更多结构化的评估机制，而不仅仅是用户评分。作为未来的工作，我们提议整合一个评估代理模型，这是一个系统内部模块，在将 AI 生成的响应呈现给用户之前检查其准确性、相关性和完整性。基于我们的经验，仅靠用户反馈往往不足以发现事实错误或不完整的回应，特别是在那些缺少或误导性信息可能产生严重后果的领域中。一个自动评估代理可以在检测到弱点时触发第二阶段检索或提示重新制定，从而创建自适应反馈循环。我们相信这样的机制对于提高 RAG 系统在高风险、实际应用中的可靠性和可信度是至关重要的。

参考文献

[1] A. Nguyen-Duc, B. Cabrero-Daniel, A. Przybylek, C. Arora, D. Khanna, T. Herda, U. Rafiq, J. Melegati, E. Guerra, K.-K. Kemell, M. Saari, Z. Zhang, H. Le, T. Quan, and P. Abrahamsson, "Generative artificial intelligence for software engineering – a research agenda," 2023. [Online]. Available: https://arxiv.org/abs/2310.18648

[2] A. Xu, T. Yu, M. Du, P. Gundecha, Y. Guo, X. Zhu, M. Wang, P. Li, and X. Chen, "Generative ai and retrieval-augmented generation (rag) systems for enterprise," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, ser. CIKM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 5599 – 5602. [Online]. Available: https://doi.org/10.1145/3627673.3680117

[3] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, and Q. Li, "A survey on rag meeting llms: Towards retrieval-augmented large language models," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 6491 – 6501. [Online]. Available: https://doi.org/10.1145/3637528.3671470

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2021. [Online]. Available: https://arxiv.org/abs/2005.11401

[5] S. Gupta, R. Ranjan, and S. N. Singh, "A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions," 2024. [Online]. Available: https://arxiv.org/abs/2410.12837

[6] N. Chirkova, D. Rau, H. Déjean, T. Formal, S. Clinchant, and V. Nikoulina, "Retrieval-augmented generation in multilingual settings," in *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, S. Li, M. Li, M. J. Zhang, E. Choi, M. Geva, P. Hase, and H. Ji, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 177–188. [Online]. Available: https://aclanthology.org/2024.knowllm-1.15/

[7] R. D. Pesl, J. G. Mathew, M. Mecella, and M. Aiello, "Advanced system integration: Analyzing openapi chunking for retrieval-augmented generation," 2024. [Online]. Available: https://arxiv.org/abs/2411.19804

[8] R. T. de Lima, S. Gupta, C. Berrospi, L. Mishra, M. Dolfi, P. Staar, and P. Vagenas, "Know your rag: Dataset taxonomy and generation strategies for evaluating rag systems," 2024. [Online]. Available: https://arxiv.org/abs/2411.19710

[9] S. Yu, C. Tang, B. Xu, J. Cui, J. Ran, Y. Yan, Z. Liu, S. Wang, X. Han, Z. Liu, and M. Sun, "Visrag: Vision-based retrieval-augmented generation on multi-modality documents," 2024. [Online]. Available: https://arxiv.org/abs/2410.10594

[10] J. Zhang, Q. Zhang, B. Wang, L. Ouyang, Z. Wen, Y. Li, K.-H. Chow, C. He, and W. Zhang, "Ocr hinders rag: Evaluating the cascading impact of ocr on retrieval-augmented generation," 2024. [Online]. Available: https://arxiv.org/abs/2412.02592

[11] J. Chen, D. Xu, J. Fei, C.-M. Feng, and M. Elhoseiny, "Document haystacks: Vision-language reasoning over piles of 1000+ documents," 2024. [Online]. Available: https://arxiv.org/abs/2411.16740

[12] S. Zhao, Y. Huang, J. Song, Z. Wang, C. Wan, and L. Ma, "Towards understanding retrieval accuracy and prompt quality in rag systems," 2024. [Online]. Available: https://arxiv.org/abs/2411.19463

[13] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, and B. Cui, "Retrieval-augmented generation for ai-generated content: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2402.19473

[14] D. Abrahamyan and F. H. Fard, " StackRAG Agent: Improving Developer Answers with Retrieval-Augmented Generation ," in *2024 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2024, pp. 893–897. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICSME58944.2024.00098

[15] M. Ahmed, M. Dorrah, A. Ashraf, Y. Adel, A. Elatrozy, B. E. Mohamed, and W. Gomaa, "Codeqa: Advanced programming question-answering using llm agent and rag," in *2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, 2024, pp. 494–499.

[16] L. Shi, M. Kazda, B. Sears, N. Shropshire, and R. Puri, "Ask-eda: A design assistant empowered by llm, hybrid rag and abbreviation de-hallucination," in *2024 IEEE LLM Aided Design Workshop (LAD)*, 2024, pp. 1–5.

[17] A. A. Khan, M. T. Hasan, K. K. Kemell, J. Rasku, and P. Abrahamsson, "Developing retrieval augmented generation (rag) based llm systems from pdfs: An experience report," 2024. [Online]. Available: https://arxiv.org/abs/2410.15944

[18] X. Wang, Z. Wang, X. Gao, F. Zhang, Y. Wu, Z. Xu, T. Shi, Z. Wang, S. Li, Q. Qian, R. Yin, C. Lv, X. Zheng, and X. Huang, "Searching for best practices in retrieval-augmented generation," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 17 716–17 736. [Online]. Available: https://aclanthology.org/2024.emnlp-main.981/

[19] Y. B. Sree, A. Sathvik, D. S. Hema Akshit, O. Kumar, and B. S. Pranav Rao, "Retrieval-augmented generation based large language model chatbot for improving diagnosis for physical and mental health," in *2024 6th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*, 2024, pp. 1–8.

[20] A. Naeem, T. Li, H.-R. Liao, J. Xu, A. M. Mathew, Z. Zhu, Z. Tan, A. K. Jaiswal, R. A. Salibian, Z. Hu, T. Chen, and Y. Ding, "Path-rag: Knowledge-guided key region retrieval for open-ended pathology visual question answering," 2024. [Online]. Available: https://arxiv.org/abs/2411.17073

[21] H. M. T. Alam, D. Srivastav, M. A. Kadir, and D. Sonntag, "Towards interpretable radiology report generation via concept bottlenecks using a multi-agentic rag," 2025. [Online]. Available: https://arxiv.org/abs/2412.16086

[22] Z. Guo, L. Xia, Y. Yu, T. Ao, and C. Huang, "Lightrag: Simple and fast retrieval-augmented generation," 2024. [Online]. Available: https://arxiv.org/abs/2410.05779

[23] G. Gamage, N. Mills, D. De Silva, M. Manic, H. Moraliyage, A. Jennings, and D. Alahakoon, "Multi-agent rag chatbot architecture for decision support in net-zero emission energy systems," in *2024 IEEE International Conference on Industrial Technology (ICIT)*, 2024, pp. 1–6.

[24] B. Sarmah, D. Mehta, B. Hall, R. Rao, S. Patel, and S. Pasquali, "Hybridrag: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction," in *Proceedings of the 5th ACM International Conference on AI in Finance*, ser. ICAIF '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 608 – 616. [Online]. Available: https://doi.org/10.1145/3677052.3698671

[25] J. Jang and W.-S. Li, "Au-rag: Agent-based universal retrieval augmented generation," in *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, ser. SIGIR-AP 2024. New York, NY, USA: Association for Computing Machinery, 2024, p. 2 – 11. [Online]. Available: https://doi.org/10.1145/3673791.3698416

[26] S. Zeng, J. Zhang, B. Li, Y. Lin, T. Zheng, D. Everaert, H. Lu, H. Liu, H. Liu, Y. Xing, M. X. Cheng, and J. Tang, "Towards knowledge checking in retrieval-augmented generation: A representation perspective," 2024. [Online]. Available: https://arxiv.org/abs/2411.14572

[27] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek, "Seven failure points when engineering a retrieval augmented generation system," in *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*, ser. CAIN '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 194 – 199. [Online]. Available: https://doi.org/10.1145/3644815.3644945

[28] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J. Wen, "A survey on large language model based autonomous agents," 12 2024.

[29] X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai, Y. Gu, H. Ding, K. Men, K. Yang, S. Zhang, X. Deng, A. Zeng, Z. Du, C. Zhang, S. Shen, T. Zhang, Y. Su, H. Sun, M. Huang, Y. Dong, and J. Tang, "Agentbench: Evaluating LLMs as agents," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=zAdUB0aCTQ

[30] A. Zeng, M. Liu, R. Lu, B. Wang, X. Liu, Y. Dong, and J. Tang, "AgentTuning: Enabling generalized agent abilities for LLMs," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 3053–3077. [Online]. Available: https://aclanthology.org/2024.findings-acl.181/

[31] A. Singh, A. Ehtesham, S. Kumar, and T. T. Khoei, "Agentic retrieval-augmented generation: A survey on agentic rag," 2025. [Online]. Available: https://arxiv.org/abs/2501.09136

[32] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, "Corrective retrieval augmented generation," 2024. [Online]. Available: https://arxiv.org/abs/2401.15884

[33] S. Li, L. Stenzel, C. Eickhoff, and S. A. Bahrainian, "Enhancing retrieval-augmented generation: A study of best practices," in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 6705–6717. [Online]. Available: https://aclanthology.org/2025.coling-main.449/

[34] S. Krishna, K. Krishna, A. Mohananey, S. Schwarcz, A. Stambler, S. Upadhyay, and M. Faruqui, "Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation," 2025. [Online]. Available: https://arxiv.org/abs/2409.12941