© 2026年 IEEE。允许个人使用本材料。对于所有其他用途，无论是在当前还是未来的任何媒体中，均须获得 IEEE 的许可，包括为广告或宣传目的重印/再版本材料、创作新的合集作品、转售或重新分发至服务器或列表，或在其他作品中重复使用本作品中受版权保护的任何组件。**正式版本可在 IEEE Xplore 获取（DOI：10.1109/ICKG70389.2026.00009，即将解析）**。

SuSi-RAG：面向设备维护的混合图检索增强生成

Hansi Zhang†罗伯特·博世有限公司上海，中国
hans.zhang@cn.bosch.com

Wilma Johanna Schmidt†罗伯特·博世有限公司伦宁根，德国柏林自由大学¨柏林柏林，德国
wilma.schmidt@bosch.com

Wanlin Yang罗伯特·博世有限公司上海，中国复旦大学上海，中国
wanlin.yang@cn.bosch.com

Xiaozhi Shen罗伯特·博世有限公司苏州，中国
xiaozhi.shen@cn.bosch.com

Adrian Paschke弗劳恩霍夫开放通信系统研究所德国柏林柏林自由大学¨柏林德国柏林
adrian.paschke@fokus.fraunhofer.de

Evgeny Kharlamov罗伯特·博世有限公司德国伦宁根奥斯陆大学挪威奥斯陆
evgeny.kharlamov@de.bosch.com

arXiv:2611.30588v1 [cs.IR] 23 Nov 2026

摘要—**在制造领域，设备维护任务依赖专家知识。相关文档通常以异构格式存储于不同来源。利用**大语言模型 **结合**检索增强生成（***RAG***）**来整合分散的知识颇具前景，但传统的基于向量的检索增强生成在应对真实维护查询中常见的两种特征时存在不足：复杂因果链和跨领域综合。为此，我们提出 *SuSi-RAG*，一种混合图检索增强生成解决方案，通过子图相似度匹配用户查询与每条维护记录。本研究展示了一个**实际运行中的**系统：*SuSi-RAG*已在博世生产线上持续运行八 个多月，处理了*10,664* 次问答交互，答案准确率上限达到*96.25%*（已确认错误率为3.75%）。在概念验证中，与基于向量的基线系统相比，*SuSi-RAG*将top-*1* 检索准确率从61.5%提升至96.2%，并将专家评定的答案质量从44.9%提升至84.1%。我们介绍了一个集成式运营反馈闭环，用于更新*SuSi-RAG*。**

索引词—**混合图检索增强生成、知识图谱、智能维护、工业4.0**

I. 引言

十多年的工业4.00 [1] 为制造业 [2],带来了诸多益处，但也带来了日益复杂的车间设备和工业机械，其软硬件相互连接。对制造企业 [3]，而言，高效且高质量地维护这些设备在经济上至关重要；然而，所需的专业知识十分稀缺，因为经验丰富的工程师拥有宝贵的专家知识，而经验不足的人员在执行困难任务时需要支持。一个典型的维护任务由以下问题触发：

ST1 工位持续出现产品焊接废品。即使更换了电极，问题仍然存在。故障排查过程接下来该采取哪些步骤?

基于关键词的检索在此类情况下往往会失败，因为它可能检索出与电极维护相关的文档，尽管这一步骤此前已经尝试过但未成功。要解决这一问题，需要采用更复杂的方法来追踪因果链，例如从焊接废料 到位置偏移 ，再到夹爪状态 ，最后到气缸故障，并协调机械、过程和控制领域之间的信息。**复杂因果链（C1）**和**跨领域综合（C2）**是博世等大型制造企业面临的两项常见挑战。我们在图1中展示了这些挑战。语义技术，例如本体—

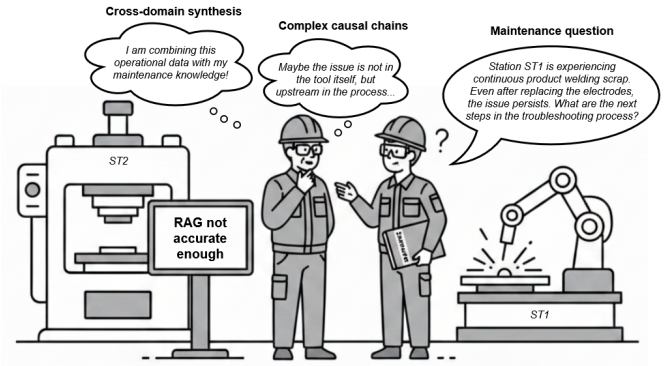

图 1. 本工作的动机：获取与设备维护相关的知识对于解决故障至关重要，但这一过程既繁琐又不简单。

方法和知识图谱（知识图谱）已在工业应用中证明了其有效性 [4]–[8]。图结构支持对因果链进行多跳搜索（**C1**），并可借助本体和知识图谱将维护知识解析为图结构。然而，在实际应用中获取这些知识既繁琐又不简单，因为目前还没有足够准确的解决方案能够满足

本研究部分得到了欧盟项目 SMARTY（GA 101140087）的支持。† 作者对本研究贡献相同。

我们的场景。一种颇具前景的方向是采用由图组件和基于大语言模型的组件构成的混合架构（例如，[9], [10]）。因此，我们提出以下研究问题：

我们如何在基于检索增强生成的解决方案中结合运行数据与领域知识，以回答真实的维护查询？这类解决方案在生产应用中是否确实有帮助？

不同于现有的知识图谱 集成到检索增强生成 系统中的方法，我们提出了*SuSi-RAG*，采用一种面向领域的检索方法来解决这一问题。用户查询和维护文档均表示为维护图，通过结合专家加权实体类型的相似度分数，以及对不完整查询进行比例协调，选出最相关的记录。与两种常见方法相结合的平衡步骤确保了稳健性，并选出最终的输出候选项。不同于大多数相似研究，我们这项面向行业和应用的工作的评估分为两部分：受控的概念验证（*PoC*）和对真实生产线的实际使用 分析。我们的贡献如下：

- **一种子图相似度检索机制**，用于维护事件，结合专家加权实体、仅对部分查询进行协调，以及将模式处理与文本处理分开，以提高检索项的搜索相关性。
- **对已部署解决方案的实际使用分析**：在博世一条选定的活跃生产线上，历时八个月，共有*10,664* 次交互，并统计了经专家确认的错误率。此外，文中还介绍了一种反馈循环机制。
- **实践洞见**：关于确保可复现性的参数配置，以及部署过程中总结的经验，以获得可靠的车间结果。
- **将*SuSi-RAG*** 与基于向量的检索增强生成进行定量比较，结果显示*SuSi-RAG*的性能几乎翻倍：检索准确率从 61.5% 提高到 96.2%，专家评定的答案质量从 44.9% 提高到 84.1%。

本文其余部分安排如下。接下来，我们将在第二节讨论相关研究，并指出针对本文场景的研究空白。第三节介绍博世的智能设备维护用例，并通过示例性维护问题说明两个主要挑战。第四节阐释采用混合图检索增强生成设计的必要性，随后详细介绍本体、知识图谱和*SuSi-RAG* 的各个模块。第五节给出配置设置，并通过两种方式评估*SuSi-RAG*：定量概念验证分析，以及对该解决方案进行持续的实际使用分析。我们将在本节末讨论方法的局限性。第六节总结经验，最后在第七节结束本文，并展望未来工作。

II. 相关工作

接下来，我们将讨论当前关于相关检索增强生成 变体以及知识图谱 在制造领域应用的研究。

基于向量的检索增强生成。增强大语言模型的一种主要策略是检索增强生成（RAG） [11]，，它将基于检索的方法与生成模型（例如，大语言模型） [12]相结合，从而提高大语言模型的事实准确性、透明度和可更新性。领域专用数据及其更新可以集成到检索组件中。然而，基于向量的检索在上下文感知能力方面有限 [13]，，不足以处理复杂因果链（**C1**）。

图检索增强生成。为解决这一问题，图检索增强生成（ [10], [13], [14]）利用了图的结构化特性。早期研究探索了结合子图检索的检索增强生成，例如*SubgraphRAG*，通过调整图的规模来适应“查询需求和下游大语言模型的能力” [15]。另一项研究采用分而治之 策略来搜索最优子图 [9]。我们的方法有所不同：它为固定的子图模式中的实体指定专家权重，分别处理基于模式和基于文本的实体值，并通过惩罚因子平衡关键词较少的查询的相似度分数。*SuSi-RAG* 将子图相似度排序与另外两种方法相结合，以提高方法的鲁棒性，详见第四节。

混合检索增强生成。混合检索增强生成框架结合检索增强生成与图检索增强生成方法，以兼具两种架构的优势 [16]。Xu等 [17] 比较了将知识图谱用于知识检索的不同策略。鉴于长文本输入带来的挑战，作者提出了一种基于知识图谱的双通路方法，以“提升检索质量，并生成更准确、连贯的输出” [17]。他们通过“显式比较知识图谱与从查询中提取的实体关系” [17]来改进提示。作者指出，跨领域查询以及大规模数据下的查询性能都面临巨大挑战。我们的工作尤其着力解决跨领域这一挑战。

工业知识图谱与检索增强生成设置。知识图谱已被证明能够整合制造数据 [5], [7]。近年来，制造业中的检索增强生成设置也日益受到关注 [18], [19]。

总而言之，据我们所知，此前尚无研究提出一种混合图检索增强生成方法，将专家加权的维护事件子图相似度匹配算法与向量存储相结合，并对其实际应用进行评估。除了针对加权实体的具体细节，以及对仅有部分匹配的查询所采用的惩罚因子外，*SuSi-RAG* 还在真实生产条件下评估了多个相关特征。

III. 智能设备维护用例与需求

我们介绍目标解决方案的用例与需求。

A. 用例描述

维护描述了“监测并使工厂、设备和设施保持正常运行的功能” [20]。本文聚焦于设备维护。维护人员分析运行数据，以识别潜在原因并执行有效的维护。一项显著挑战在于，依赖宝贵且丰富维护经验的资深工程师与可能难以进行根因识别和故障排查的初级同事之间存在知识鸿沟。为弥合这一鸿沟，维护人员遵循标准程序；这些程序经过记录，并融入了资深工程师提供的复杂实用见解。虽然维护

表 I T两类典型设备维护问题。说明：* 答案由维护专家组成的利益相关方小组提供；挑战：挑战，*C1*：复杂因果链；*C2*：跨领域综合。

Ch.	问题（简略）	标准答案（简略）*
C1	站点 *ST1* 出现持续产生产品焊接废品。即使更换了电极，问题仍然持续存在。接下来怎么办?	接下来的故障排查步骤包括：1. 位置观察：观察焊接位置的物理偏移。2. 检查关位置和电机销校正-气缸。3. 根因定位夹爪 *G1*、*G2*处的根因定位。如果它们无法拧紧，则进行更换。4. 更换后，监控传感器信号状态。
C2	站点 *ST2* 报告持续的测试失败；压力机无法复位（超时 *T1*，错误 *E1*）。如何处理?	1. 检查电源引脚状态。2. 如果压力机未能到达原位，退回使用调试软件 *ABC*。3. 检查并修改压机限位设置中的过程控制参数；比较类似工位（例如 *ST3*），以优化至 a 更稳定的范围。

文档可以为初级员工提供大量指导，同时也是持续改进维护规程的基础。博世此前部署的一种检索增强生成解决方案 [21]显示出准确性方面的局限性，我们基于多次测试运行对其进行了量化：平均检索准确率仅为*61.5%*，这限制了该解决方案识别最相关源文档的能力。生成准确率甚至更低，仅为*44.9%*，导致对查询给出不完整或不正确的答案。表 I 展示了两个典型的设备维护问题及专家标准答案。问题 1 展示了具有复杂关系和实体的多跳因果链查询（**C1**）。多跳查询是指需要跨异构文档关联碎片化知识的查询。这类查询在设备维护任务中十分典型，但也存在于医疗保健（参见例如 [22]）或金融（参见例如 [23]）等其他领域。问题 2 是跨领域综合的一个示例：将运行数据与维护程序指南和潜在事件原因中的领域知识联系起来，并进一步将机械领域的要素与工艺参数进行综合（**C2**）。

B. 目标解决方案的需求

我们根据上述挑战和专家反馈，提炼出四项具体需求：**准确性（R1）：** 答案准确率 ≥ 0.7 是该用例具有价值的必要条件。**透明性（R2）：** 将答案关联到源文档，以便验证回答并增强用户信任。**持续更新（R3）：** 知识必须能够根据新的事件轻松更新，使答案反映当前状态。**可扩展性（R4）：** 面对大型数据语料库时，应能快速生成答案。该解决方案还应易于部署。

IV. *SuSi-RAG*用于智能设备维护

接下来，我们解释为何需要混合架构，然后介绍知识图谱、提及数据摄取管道，并进一步介绍所有运行时模块，见图2。

A. 混合图检索增强生成的动机

大语言模型提供了用户友好的自然语言接口，但缺乏专有的（维护）知识。检索增强生成的优势 [11]在于提高准确性（**R1**）和透明度（**R2**）、保持时效性（**R3**），并提升可扩展性（**R4**）。然而，纯粹基于向量的检索增强生成方案在上下文感知能力有限、检索过程中存在噪声，以及信息相互矛盾或不符合事实（ [13], [24]）等方面存在不足。图检索增强生成会考虑数据内部的关系和相互联系，从而实现更准确、全面的信息检索，尤其适用于特定领域的场景。图检索增强生成擅长处理复杂因果链（**C1**），但在处理近期的非结构化信息（例如尚未纳入知识图谱的最新维护事件）或详细的文本说明时存在局限。因此，我们采用混合检索增强生成（参见例如 [16], [17]），结合基于向量和基于图的检索。

考虑表 I，图检索增强生成可以识别**C1** 中的气缸故障 为根本原因，并提供诊断路径以提高对用户的透明度。针对气缸故障维修 进行额外的向量搜索，可以提供基于文本的维修说明。图检索增强生成能够识别并关联实体（例如，机械症状 与通过工位类比修改工艺参数），并检索**C2**中的相似站点；而基于向量的检索增强生成则会检索包含这些实体相关说明的文档。*SuSi-RAG* 通过整合历史维护知识，为工程师的日常维护任务提供支持，并针对所选博世生产线上的机器错误提供精准的解决建议。通过对知识图谱模块的专门整合，**SuSi-RAG**能够有效执行相似性搜索，纳入语义含义和复杂关系，同时减少检索信息中的冗余和噪声。

B. 本体与知识图谱

本体开发和半自动化知识图谱构建流程已在先前的工作 [21] 中展示。该本体是对*BoschI4.0*知识图谱 [25]中本体集合的维护扩展，后者复用了制造领域中的大多数重要概念，例如机器、组件、工厂、线路等[1]。图 3 展示了该本体及其指标，其中核心类 MaintenanceEvent 与 MachineComponent 和 ErrorCode 等类相关联。

C. 数据摄取管道与嵌入

设备维护文档会直接嵌入向量数据库，构建完成的知识图谱则单独进行向量化。在运行时，我们的解决方案会访问这两个数据库。基于嵌入的检索是我们进行文档搜索的方法基础。我们使用 面向检索增强生成的双语与跨语言嵌入（*BCEmbedding*）模型，这是一种专为检索增强生成实现设计的稳健高效嵌入框架，具有出色的性能 [26]。该模型针对多语言和领域特定文本进行了优化，

[1]https://github.com/eclipse-esmf/esmf-manufacturing-information-model

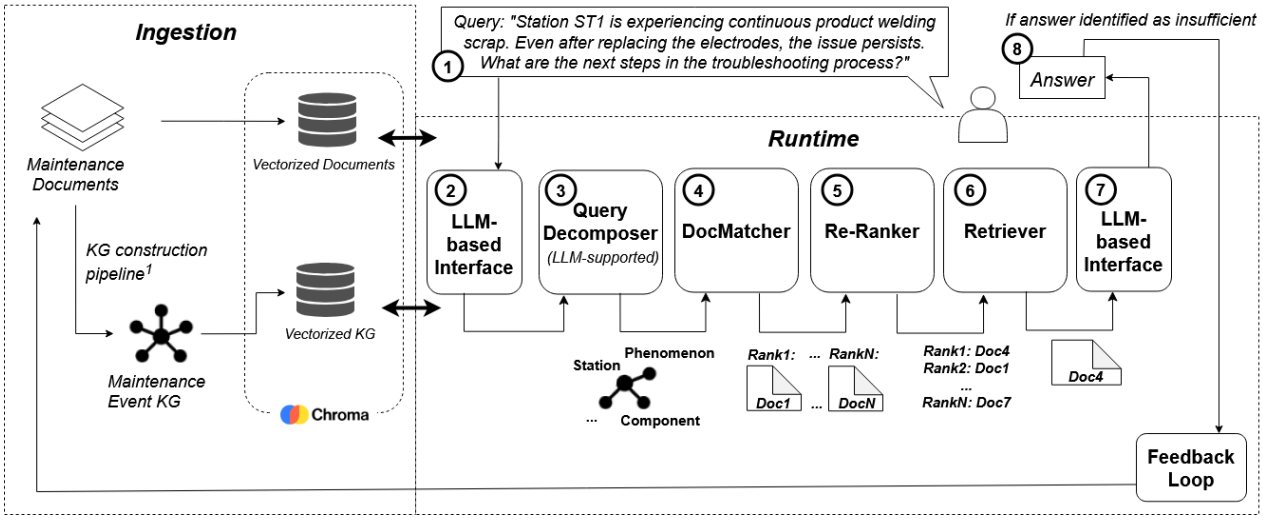

图 2. *SuSi-RAG* 的数据摄取与运行时概览。

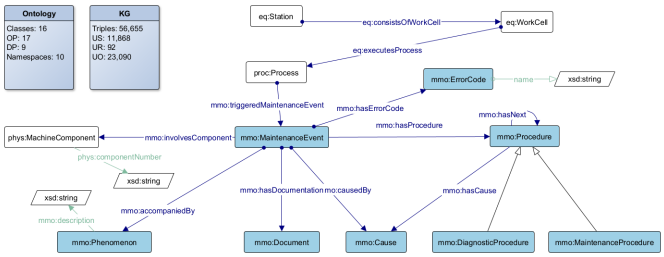

图 3. 本体 **MMO** [21] 的主要概念，以及 **MMO** 和生成的维护知识图谱的指标。图例：OP：对象属性；DP：数据类型属性；US：唯一主语；UR：唯一关系；UO：唯一宾语。

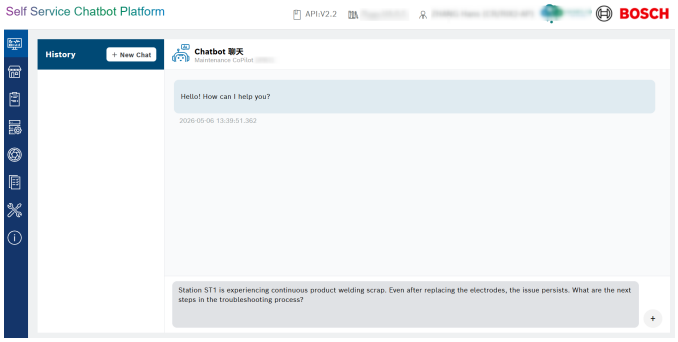

图 4. *SuSi-RAG* 的基于大语言模型的接口 运行时截图。

因此非常适合我们的用例。该方法包含一个 嵌入模型和一个重排序模型，二者协同工作，以提升检索的精确度和准确性。

D. 运行时模块

接下来，我们将介绍查询在各个模块中的处理过程，并逐一说明各模块，见图2。**基于大语言模型的接口**。 首先，用户（例如维护工程师）在基于大语言模型的接口中输入问题。这个自然语言接口是一个聊天机器人，如图4所示。**查询分解器**。 查询分解器 模块接收用户问题作为输入，首先通过大语言模型代理进行零样本判断，确定问题是否与维护问题相关（“是”/“否”）；如果相关，则提取预定义的维护相关键值对。在提取过程中，键k_i、$i = 1..n$表示本体中与维护事件关联的节点，而值v_i 则直接从问题中提取。如果问题中未提及键k_i ，其对应的值v_i 则为“None”。例如，给定问题Q =“站点 *S1* 发出警报，组件 *C2*：轴停止。碰撞。该如何处理？”，我们提取的键包括站点、现象、过程、机器部件和错误代码，对应的结果如下：

如下：”站点”:”站点 *S1*”，”现象描述”: ”轴停止。碰撞。”，”过程”: *None*，”机器部件”:”组件 *C2*”，”错误代码”: *None*。本体中的更多键（节点）如图3所示。如图2所示，提取出的键值对可以表示为维护事件子图，供后续模块进行比对和搜索。对于与先前查询高度相似的查询，我们会复用存储在向量数据库中的历史查询分解器结果，并跳过查询分解过程，从而显著加快预检索模块的处理速度。

文档匹配器。 图2展示了文档匹配器 模块的输入是一个经过解析的问题Q ，其表示为维护子图；此外，每个文档E_i 均表示为维护事件子图。这些子图包含维护事件、站点、错误代码、事件原因等节点。重要元素是节点值。为识别最相关的文档，我们开发了子图相似度分数SE_i ，以实现针对维护事件的细粒度相似度匹配。具体而言，我们使用 *BCEmbedding*模型生成节点值的向量化表示。首先，我们计算分解后的查询中的每个键与每个维护事件子图中对应实体之间的匹配分数。对于

文本实体（如现象），我们使用余弦相似度。对于基于模式的实体（例如机器部件），由于其遵循由数字和字母组成的特定模式，我们采用正则表达式并使用二元评分：匹配（*1*）或不匹配（*0*）。为考虑不同实体类型的重要性差异，我们根据专家建议为每种实体类型分配动态加权因子。由于某些维护文档可能缺少某些实体或其值为空，从而导致最终匹配分数存在差异，因此我们通过将每个实体的权重相对于文档中所有已提取实体的总权重进行缩放来对权重进行归一化。无论可用实体数量如何变化，这都能确保匹配分数保持一致且具有可比性。

每个用户查询Q 都可以分解为一组键值对（k_j,v_j）。然后，根据专家对实体类型的权重分配，我们将实体k_j 的权重定义为w_{k_j}。因此，用户查询Q 与维护事件E_i 之间的总匹配分数S_{E_i} 计算如下：

$$S_{E_i} = \sum_{(k_j,v_j)\in Q} \Big[\Big(I(\text{type}(k_j) = \text{pattern}) \cdot I(E_i.\text{sec}[\text{value}] = v_j) \cdot \text{norm}(E_i, w_{k_j}) + I(\text{type}(k_j) = \text{text}) \cdot \text{sim}(E_i.\text{sec}[\text{value}], v_j) \cdot \text{norm}(E_i, w_{k_j}) \Big) \Big] \tag{1}$$

其中，$E_i.sec$ 表示E_i中的一个关键元素，$E_i.sec[\text{value}]$ 表示实体$E_i.sec$的值，$\text{norm}(E_i,w_{k_j})$表示E_i 中实体k_j 的归一化权重，计算如下：

$$\text{norm}(E_i, w_{k_j}) := \frac{w_{k_j}}{\sum_{k_l \in E_i, k_l \neq None} w_{k_l}},$$

其中，$I(\text{condition})$是一个布尔指示函数：若括号内的条件为真，则值为1，否则为0；$\text{sim}(E_i.sec[\text{text}],v_j)$表示文本节点的余弦相似度[2]。

为防止仅匹配到一个实体的文档获得过高的匹配分数，我们使用惩罚因子，根据匹配实体类型所占的比例对原始匹配分数S_{E_i} 进行缩放。我们将此惩罚因子记为PF，并将最终缩放后的得分S'_{E_i} 计算如下：

$$S'_{E_i} = S_{E_i} \cdot \frac{1}{1 + \left(\frac{1}{|matched_sec(E_i)|} - \frac{1}{|all_sec|}\right)}, \tag{2}$$

其中，S_{E_i} 是按照公式（1）计算得出的维护事件E_i 的原始匹配分数，$|matched\ sec(E_i)|$是维护事件E_i 中匹配实体的数量（即查询Q中存在这些实体的值），而 $|all\ sec|$是E_i中所有实体的数量。

为了高效地计算所需输出，我们遵循算法1中描述的流程，并进一步将向量化知识图谱节点及其值存储在 Chroma 中[3]，以加快检索速度。

重排序器和检索器。 重排序器模块通过对排名靠前的k 个文档重新评分来优化结果。为此，我们使用*BM25*[4] 匹配方法和*BCE-RerankerModel*，重新计算用户查询与各文 *BCE-RerankerModel* 档之间的相似度。该模型会联合编码

[2]参见，例如：https://en.wikipedia.org/wiki/Cosine similarity [3]https://docs.trychroma.com/ [4] https://pypi.org/project/BM25/

算法1文档匹配器

输入： 分解后的查询Q、维护事件集 E

输出： 文档匹配分数列表S_E

1：为每个维护事件初始化得分 $S_{E_i} \leftarrow 0$

2: **对于**每个$E_i \in E$，在E **中执行**

3: **对于**每个 (k_j,v_j)，在Q**中执行**

4: **如果**$\text{type}(k_j) =$ 模式 **则**

5: 检查$E_i.sec = k_j$是否

6: $s \leftarrow I(E_i.sec[\text{value}] = v_j)$ {精确匹配基于模式的实体}

7: **否则如果** $\texttt{type}(k_j) = \texttt{textual}$ **则**

8: 检查 $E_i.sec = k_j$ 是否

9: $s \leftarrow \text{sim}(\text{embed}(E_i.sec[\text{文本}]),\text{embed}(v_j))$

10: **结束 if**

11: $s \leftarrow s \times$ 范数（E_i，w^{k_j}）

12: $S_{E_i} \leftarrow S_{E_i} + s$

13: **结束循环**

14: $S'_{E_i} \leftarrow S_{E_i} \times PF$

15: 追加 S'_{E_i} 到S_E

16: **结束循环**

17：**返回**文档匹配分数列表S_E

查询和每篇文档，将其作为单个序列处理，因此比双编码器能产生更丰富的词元级交互。通过倒数排名融合（*RRF*）方法，我们整合基于子图的相似度分数、*BM25*（其权重略高于其他方法）以及*BCE* 重排序模型的相似度分数，将三种方法的排名融合为一个稳健的最终排名。检索器根据重排序器的结果，从知识图谱中提取相应的子图（即维护事件）。相关性分数较高的子图被认为与用户查询的关联更紧密。使用*SPARQL* 查询模板，我们从知识图谱中提取有关所选维护事件的所有相关信息，并以*RDF* 数据形式呈现。随后通过后处理步骤对*RDF* 数据进行压缩、简化并按逻辑顺序排列；这些数据往往包含冗余信息或冗长的统一资源标识符。后处理后的数据简洁且结构清晰，便于大语言模型高效使用。

反馈循环。可选的反馈循环 为用户提供了两种标记答案不充分的方式：将其标记为知识缺口或系统的错误映射问题。这两种标记都会触发手动任务，具体说明见第五节。

V. 评估

我们开展了双重评估：首先进行受控的定量概念验证 分析，以建立基线性能；随后在博世生产线上进行持续分析，历时八个月。

A. 可复现性与设置

表 II 展示了采用大语言模型 *PoC GPT4.5* 的配置，其中通过 *BCEmbedding* 实现嵌入和重排序，使用 Chroma 作为向量存储，并通过 SPARQL 查询访问知识图谱。

表 II SUSI-RAG概念验证 实现配置。

组件	配置
查询分解与意图	*GPT 4.5*
生成式大语言模型	*GPT 4.5*
大语言模型温度	*0.1*
嵌入	*BCEmbedding*
向量存储	*Chroma*
文档分块大小	*500* 个字符
文档分块重叠	*100* 个字符
实体权重（由专家确定）	现象描述：0.40，组件：0.15，错误代码：0.15，站点：0.10，过程：0.10，线路：0.10
源的 RRF 常数	子图：45，BM25 重排序：60，BCE 重排序器：40

B. 概念验证评估

我们基于 26 个由专家整理的维护查询对 *PoC* 进行评估。这些查询源自经专家验证的源文档及其参考答案，涵盖六个生产站点和多种故障类型，包括传感器、通信、运动、机械、螺钉处理和软件故障。检索性能衡量系统在前 k 个结果中识别正确源文档的能力。表 III 显示，*SuSi-RAG* 的 Top-1 得分为 96.2%（26 个查询中答对 25 个；95% Wilson 置信区间（CI）：81.1%–99.3%），优于一种包含重排序、通过 *Azure AI Search*实现的基于向量的 RAG 基线，其得分为 61.5%（26 个查询中答对 16 个；95% Wilson CI：42.5%–77.6%）。*SuSi-RAG*较高的 Top-1 得分对于降低车间人员在时间敏感型维护任务中的认知负荷至关重要。在 26 个查询中，*SuSi-RAG* 额外正确回答了 9 个基线检索增强生成失败的案例，同时没有引入新的失败案例。部分消融实验，即仅移除重排序组件（*SuSi-RAG*不含重排序，18/26；95% Wilson CI：50.0%–83.5%），显示出基于图的检索相较于基线（含重排序）的增益，以及重排序组件在 *SuSi-RAG* 中带来的明显增益。需要注意的是，基线与不含重排序的 *SuSi-RAG* 的置信区间存在重叠，因此在 26 个查询的规模下，这一差异只能作为指示性结论。这表明结构化检索（例如通过基于图的检索和专家权重）与交叉编码器重排序可能具有互补作用：前者用于改进候选集，后者用于将正确文档推至第 1 位。其他组件（专家权重、惩罚因子、BM25）未进行单独消融。完整的消融研究将作为未来工作。

我们根据专家正确性评分（答案质量）、完整性以及与专家答案的余弦相似度，对生成的答案进行了评估。领域专家对照标准答案，手动给出专家正确性评分：0（错误）、1、2 或 3（完全正确）。报告的百分比为评分总和除以 23 个查询的最高总分（$69 = 3 * 23$）。在 26 个查询中，有 3 个必须排除：其中一个查询的基线答案因疏忽未由专家评估，且无法让同一批专家重新进行实验；另外两个查询是在专家评估后加入的，同样无法让同一批专家重新进行实验。专家评估后，又有两个查询被加入集合，但同样无法让同一批专家重新进行实验。两个系统的专家正确性评分均基于相同的 23 个查询进行评估，其他所有结果则基于 26 个查询得出。完整性是根据生成答案中与标准答案里由专家识别的关键要素相匹配的内容所占比例计算的。余弦相似度衡量生成答案与标准答案的嵌入表示之间的相似程度。表IV 显示，*SuSi-RAG* 的专家正确性评分为 84.1%，几乎是基线检索增强生成的两倍。使用*SuSi-RAG*后，余弦相似度有所提高，完整性的提升尤为明显。

表 III 检索过程评估（以 % 计）。注：N = 表示评估查询的数量。

配置	Top-1 (n=26) 查询	Top-1 (n=26) %	Top-3 (n=26) 查询	Top-3 (n=26) %
基线检索增强生成	16/26	61.5%	24/26	92.3%
SuSi-RAG（不含重排序）	18/26	69.2%	25/26	96.2%
SuSi-RAG	**25/26**	**96.2%**	**26/26**	**100%**

表IV 答案生成评估（%）。说明：N = 为评估查询的数量。

方法	专家正确性得分 (n=23)	完整性 (n=26)	余弦相似度 (n=26)
基线检索增强生成	44.9%	38.0%	73.2%
SuSi-RAG	**84.1%**	**79.6%**	**83.2%**

C. 在用生产分析

SuSi-RAG 已在博世的一条生产线上运行了八 个月。初始部署完成后，核心开发团队将*SuSi-RAG* 移交给运营IT工厂团队，该团队开发了专用的维护聊天机器人仪表板（图 5），用于追踪关键绩效指标并生成运营洞察。该系统在

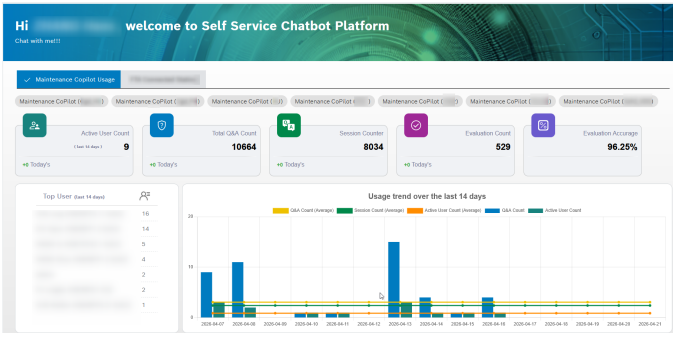

图 5. 工厂人员开发的维护聊天机器人仪表板，用于追踪真实使用情况和准确性指标。

生产环境具有以下特点：

- **使用量：**截至2026年4月，系统已成功处理*10,664* 次问答交互，涵盖*8,034*个独立会话。

- **已确认错误率：**在 10,664 个答案中，用户标记了 529 个答案供审核；维护专家确认其中 400 个答案不正确，因此已确认错误率为 3.75%（95% Wilson 置信区间：3.41–4.13%）。由于未标记的答案未经验证，96.25% 只是准确性的上限，而非经验证的数值，但它反映了用户对系统实际感知的可靠性。

反馈循环确保知识图谱持续演进。我们监控两项主要指标（见图 6）：1. **知识缺口**用于识别历史数据尚未处理、因而尚未纳入知识图谱的特定区域-线路-车站组合。这会提醒 IT 工厂团队优先为这些站点安排数据摄取。2. **错误映射问题**发生在用户提出的问题无法在知识图谱中找到相关历史维护问题时。这表明需要修正现有三元组，或从近期事件中摄取新的“标准答案”。

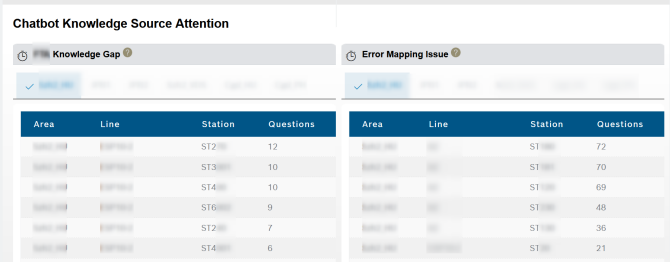

图 6. 用于识别并弥合维护知识缺口的反馈监控器。

D. 局限性

尽管我们的工作在实践中很有用，但仍存在以下局限性：

消融。 与基于向量的方法以及*SuSi-RAG*（不含重排序）进行基线比较，限制了对各个独立组件增益价值的分析。下一步最重要的是开展一项受控的改进研究，并进行消融实验。

样本量。 经过验证的概念验证集仅包含*26* 个查询，但由于时间和成本压力，收集更多专家标注数据颇具挑战。

对标记答案的专家验证。用户标记了正在使用的 529 个答案。未标记的答案未经验证，因此可能还存在更多错误。

可复现性。虽然我们展示了配置设置和参数（见表 II），但出于专有原因，可能无法公开*SuSi-RAG* 的实际数据和代码。

VI. 经验总结

我们总结了五条经验，并补充介绍了 SuSi-RAG 在博世的部署经验：系统运行了八个多月，交互次数超过 10,000 次，由此深入了解了工业4.0中语义维护系统的生命周期。

A. 通过子图匹配实现粒度与精度

系统成功的一个主要原因，是从粗粒度的文本块匹配转向细粒度的维护事件子图匹配。通过纳入领域特定的节点权重，并为部分匹配设置惩罚因子，我们在概念验证中实现了近乎完美的检索准确率。定量结果证实，对于复杂的设备维护，知识图谱提供的结构上下文比向量空间中的纯相似度更有效。一些关键流程嵌入在图像中，我们将在未来的工作中处理这一问题（参见第七节）。

B. 从静态知识转向不断演进的生态系统

为期八个月的部署凸显了持续更新工业知识图谱的必要性。我们不再采用被动式文档记录，而是利用维护聊天机器人仪表板推动持续知识生命周期：

使用驱动的形式化：通过监控知识图谱，用户需求如今决定了形式化工作的优先级。这使知识图谱从被动工具转变为战略驱动因素，将专家投入集中于最常见的信息缺口。

管理语义债务：我们使用错误映射问题来量化语义债务，即编码三元组与车间事件之间的差异。将这些问题记录为黄金答案会触发自我纠正循环，确保知识图谱演进与设备更新同步。

这一主动循环确保系统在高容量生产环境中保持弹性，并在*96.25%* 的上限答案准确率下维持稳定，覆盖*10,664*次交互。

C. 速度与容量调整（*R4*）

本部分总结了三项 旨在满足速度和容量需求的小幅改进：(i) 为实现较短的回答时间（**R4**），我们通过复用存储在向量数据库中的历史查询分解器结果，大幅提升查询分解期间的推理速度（见第四节中的查询分解器）；(ii) 由于相似性搜索可能变慢，我们通过知识图谱向量化提升速度（见第四节中的文档匹配器）；(iii) 在向大语言模型发出提示之前，我们先压缩*RDF* 数据（见第四节中的重排序器和检索器）。

D. 需求满足情况

R1 在概念验证中得到满足，专家评定的答案质量（84.1%）高于要求的 0.7。经过八个月的实际部署评估后，我们发现，96.25% 的准确率上限和 3.75% 的已确认错误率表明**R1**得到满足。经验证的准确率需要通过后续研究确认（见第五节D部分）。答案中提供的来源链接满足**R2**。尤其是，反馈循环满足**R3**。最后，使用驱动的形式化以及速度与容量调整充分满足**R4** 。

VII. 结论与未来步骤

我们介绍了*SuSi-RAG*，这是一种新型混合图检索增强生成解决方案，旨在应对博世设备维护中的关键知识获取与利用挑战。在其核心部分，*SuSi-RAG* 包含一种维护事件子图相似度匹配算法，该算法采用专家加权实体和部分匹配惩罚因子，并辅以向量检索。

SuSi-RAG 能够有效处理复杂因果链和跨领域信息所引发的问题，使专家评定的答案质量几乎翻倍，从44.9%提升至84.1%，并在*PoC*环境中将检索准确率从61.5%提高至96.2%。我们通过在博世生产线上成功部署八 个月展示了其实用价值；在此期间，*SuSi-RAG*处理了超过*10,000* 次交互，答案准确率上限达到*96.25%* （已确认错误率为3.75 %）。*SuSi-RAG* 的模块化设计包括一个基于大语言模型的接口、四 个主要的顺序式查询处理模块，以及第五个用于持续数据质量改进的模块。*SuSi-RAG* 确保提供高精度响应，并为博世内部的其他检索增强生成系统提供可复用组件方面的优势。未来，我们将着力解决系统的局限性。首先，我们将增加经专家验证的维护查询数量，以提高性能评估的统计可靠性。其次，我们将开展受控的完整消融研究，以确定增值组件的价值。第三，我们将纳入多模态数据，首先从有助于处理复杂维护任务的图像开始。第四，我们将探索智能体集成，例如用于复杂推理任务，以提出知识图谱补全建议。这些建议经专家审核后，无需事先存在实际事件即可添加到知识图谱中。第五，我们将通过添加检索路径和排名依据来增强可解释性。

致谢

作者感谢 Qiushi Cao、Tianyang Liu、Sebastian Monka 和 Irl´anGrangel-Gonz´alez 就本研究开展了富有价值的交流。

生成式人工智能声明

在准备本研究期间，作者使用 Gemini Enterprise 2.5 Pro（搭配 NanoBanana 2.5）制作图1，具体用于：生成图像的大部分内容。作者还使用 Gemini Enterprise 2.5 Pro 对论文的部分内容进行处理，具体用于：改写少量文本。使用该工具后，作者根据需要对内容进行了审阅和编辑，并对本出版物的内容承担全部责任。

参考文献

[1] H. Lasi、P. Fettke、H.-G. Kemper、T. Feld 和 M. Hoffmann，《工业4.0》，《商业与信息系统工程》，第6卷，第4期，2014年。[2] B. Lin、W. Wu 和 M. Song，《工业4.0：驱动因素及其对企业绩效的影响——对中国制造业的一项实证研究》，《运筹学年鉴》，第329卷，第1期，2023年。[3] L. Silvestri、A. Forcina、V. Introna、A. Santolamazza 和 V. Cesarotti，《通过工业4.0技术实现维护转型：系统性文献综述》，《工业中的计算机》，第123卷，2020年。[4] A. Iglesias Molina、J. A. Bernab´e Diaz、P. Deshmukh、P. P. Es- pinoza Arias、A. Ayll´on-Ben´ıtez、A. Fern´andez Izquierdo、J. M. Ponce- Bernab´e、S. P´erez、E. Ruckhaus、O. Corcho 等，《工业环境中的本体管理：巴斯夫本体治理运营模型（GOMO）》，*2022 ISMB Bio-Ontologies*，2022年。[5] M. Yahya、A. Ali、Q. Mehmood、L. Yang、J. G. Breslin 和 M. I. Ali，《面向工业4.0生产线、附带知识图谱生成的基准数据集》，《语义网》，第15卷，第2期，2024年。[6] M. Hofer、D. Obraczka、A. Saeedi、H. K¨opcke 和 E. Rahm，《知识图谱构建：现状与挑战》，《信息》，第15卷，第8期，2024年。

[7] E. G. Kalaycı、I. Grangel Gonz´alez、F. L¨osch、G. Xiao、A. ul Mehdi、E. Kharlamov 和 D. Calvanese，《利用虚拟知识图谱对博世制造数据进行语义集成》，载于《国际语义网大会》。施普林格，2020年。[8] T. R. Chhetri、S. Aghaei、A. Fensel、U. G¨ohner、S. G¨ul-Ficici 和 J. Martinez-Gil，《利用贝叶斯结构学习和知识图谱优化制造过程》，载于《计算机辅助系统理论—EUROCAST 2022—第18届国际会议，修订精选论文》，第13789卷。施普林格，2022年。 [在线]。链接：https://doi.org/10 1007 978 3 031 25312 6 70_。/ - - - - Y. Hu、Z. Lei、Z. [9] Zhang、B. Pan、C. Ling 和 L. Zhao，《Grag：图检索增强生成》，载于《计算语言学协会论文集：NAACL 2025》，2025年。[10] Z. Xu、M. J. Cruz、M. Guevara、T. Wang、M. Deshpande、X. Wang 和 Z. Li，《利用知识图谱进行检索增强生成，以实现客户服务问答》，载于《第47届国际ACM SIGIR信息检索研究与发展大会论文集》，2024年。[11] Y. Gao、Y. Xiong、X. Gao、K. Jia、J. Pan、Y. Bi、J. Sun、H. Wang 等，《面向大语言模型的检索增强生成：综述》，arXiv预印本 arXiv:2312.10997，2023年。[12] P. Lewis、E. Perez、A. Piktus、F. Petroni、V. Karpukhin、N. Goyal、H. K¨uttler、M. Lewis、W.-t. Yih、T. Rockt¨aschel 等，《面向知识密集型自然语言处理任务的检索增强生成》，《神经信息处理系统进展》，第33卷，2020年。[13] B. Peng、Y. Zhu、Y. Liu、X. Bo、H. Shi、C. Hong、Y. Zhang 和 S. Tang，《图检索增强生成：综述》，《ACM信息系统汇刊》，第44卷，第2期，2025年。[14] D. Edge、H. Trinh、N. Cheng、J. Bradley、A. Chao、A. Mody、S. Truitt、D. Metropolitansky、R. O. Ness 和 J. Larson，《从局部到全局：一种用于查询导向摘要的图检索增强生成方法》，arXiv预印本 arXiv:2404.16130，2024年。[15] M. Li、S. Miao 和 P. Li，《简单即有效：图和大语言模型在基于知识图谱的检索增强生成中的作用》，arXiv预印本 arXiv:2410.20724，2024年。[16] B. Sarmah、D. Mehta、B. Hall、R. Rao、S. Patel 和 S. Pasquali，《Hybridrag：整合知识图谱与向量检索增强生成，实现高效信息抽取》，载于《第5届ACM金融人工智能国际会议论文集》，2024年。[17] S. Xu、M. Chen 和 S. Chen，《利用知识图谱增强检索增强生成模型：双路径方法的创新实践》，载于《先进智能计算技术与应用》。施普林格·自然新加坡，2024年。[18] T. Zhou、Y. Liu、M. Kumar 和 L. Zou，《面向智能制造知识支持的轻量级查询自适应检索增强生成框架》，《制造系统杂志》，第86卷，2026年。[19] A. Chandrasekhar、J. Chan、F. Ogoke、O. Ajenifujah 和 A. Barati Farimani，《Amgpt：用于增材制造上下文查询的大语言模型》，《增材制造快报》，第11卷，2024年。[在线]。链接：https://www.sciencedirect.com/science/article/pii/S2772369024000409[20] R. Manzini、A. Regattieri、H. Pham 和 E. Ferrari，《工业系统维护》。施普林格，2010年。[21] H. Zhang、W. J. Schmidt、X. Shen、Q. Cao、S. Monka 和 A. Paschke，《面向图检索增强生成的智能维护聊天机器人的知识图谱构建》，载于《知识图谱行业规模化国际研讨会2025》，2025年。 [在线]。链接：https://ceur-ws.org/Vol-4064/SKGi-paper3.pdf[22] F. Wang、D. Shi、J. Aguilar、X. Cui、J. Jiang、L. Shen 和 M. Li，《LLM-KGMQA：基于医学领域知识图谱的大语言模型增强型多跳问答系统》，《知识与信息系统》，第67卷，第8期，2025年。[23] M. R. Karim、F. Hermsen、S. A. Chala、P. De Perthuis 和 A. Mandal，《面向反洗钱的可扩展半监督图学习技术》，IEEE Access，第12卷，2024年。[24] N. F. Liu、K. Lin、J. Hewitt、A. Paranjape、M. Bevilacqua、F. Petroni 和 P. Liang，《迷失在中间：语言模型如何利用长上下文》，《计算语言学协会汇刊》，2024年。[25] I. Grangel-Gonz´alez、F. L¨osch 和 A. ul Mehdi，《用于高效集成和访问制造数据的知识图谱》，载于《2020年第25届IEEE新兴技术与工厂自动化国际会议（ETFA）》，第1卷，2020年。[26] N. Youdao，《BCE mbedding：面向检索增强生成的双语与跨语言嵌入》，https://github.com/netease-youdao/BCEmbedding，2023年。