

AI Ethics to Requirements Practice: Building and Evaluating the HealthAI Ethics Assistant

Yutan Huang ✉ 

Monash University, Melbourne, Australia

Jingfan Chen ✉

Monash University, Melbourne, Australia

Fanyu Wang ✉ 

Monash University, Melbourne, Australia

Chetan Arora ✉ 

Monash University, Melbourne, Australia

John Grundy ✉ 

Monash University, Melbourne, Australia

Cheng Zhang ✉

BMC Medical Co., Ltd., Beijing, China

Abstract

Healthcare AI software teams face growing pressure to translate ethical principles and regulatory obligations into concrete requirements, design decisions, and assurance evidence. Day-to-day requirements engineering (RE) practice still lacks lightweight tool support for connecting abstract frameworks, such as the EU AI Act and the NIST AI Risk Management Framework, to actionable requirements work. This paper builds on a two-phase research project. In the first phase, we conducted an empirical interview study with healthcare AI practitioners spanning clinicians as end users of HealthAI systems and software engineers as developers of HealthAI software systems, which surfaced recurring concerns around transparency, accountability, and the difficulty of mapping regulatory obligations to concrete engineering tasks. Informed by those findings, in the second phase we designed, implemented, and conducted an initial evaluation of the *HealthAI Ethics Assistant*, an AI-enabled RE tool for healthcare AI software systems.

The tool supports practitioners in generating, validating, and comparing ethical requirements through a structured Create–Validate–Compare workflow. It was implemented as a full-stack web application and grounded in a structured knowledge base combining regulatory guidance (EU AI Act, NIST AI RMF) with practitioner concerns surfaced by the interview study. The work was conducted in close collaboration with an R&D engineer at a medical device manufacturer, who contributed to industrial problem framing and participated as the first practitioner evaluator in the case study reported here.

Our initial evaluation is exploratory, involving a single case study, and used the Technology Acceptance Model and the System Usability Scale. Traceable regulatory references were rated as the most valuable and trustworthy feature, highlighting the importance of explainability and evidence support in compliance-oriented requirements work. The main adoption challenge was not basic usability, but fitting the tool into existing engineering and regulatory workflows. We derive practice-oriented lessons for designing AI-enabled requirements tools in regulated domains: ground LLM outputs in explicit compliance sources, support multiple practitioner perspectives, make generated requirements reviewable rather than authoritative, and align tool use with existing assurance processes. These early results suggest that AI-enabled assistants can help bridge ethical AI principles and practical requirements engineering when designed as auditable, human-in-the-loop support tools rather than autonomous compliance solutions.

2012 ACM Subject Classification Software and its engineering → Requirements analysis

Keywords and phrases Requirements Engineering, Software Engineering, AI ethics, Healthcare, Large Language Models, Regulatory Compliance.



© Yutan Huang, Jingfan Chen, Fanyu Wang, Chetan Arora, John Grundy, and Cheng Zhang; licensed under Creative Commons License CC-BY 4.0

20th International Symposium on Empirical Software Engineering and Measurement (ESEM 2026).

Editors: Robert Feldt, Maria Paasivaara, Daniel Mendez, Stefan Wagner, and Marvin Muñoz Barón;

Article No. 87; pp. 87:1–87:13



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Digital Object Identifier 10.4230/LIPIcs.ESEM.2026.87

Category Software Engineering in Practice Track Paper

1 Introduction

Healthcare AI software systems (hereafter referred to as *HealthAISys*) development teams are under growing pressure to demonstrate that their systems meet ethical and regulatory standards such as the EU AI Act [10] and the NIST AI Risk Management Framework [22]. These frameworks set out clear expectations around transparency, accountability, and patient safety; the challenge is that these expectations are expressed at a level of abstraction that is difficult to act on directly in a software project. In practice, most engineering teams lack a clear process for translating a regulatory provision into something they can actually build or test. This gap between ethical guidance and requirements engineering (RE) practice is not new. Prior empirical work has documented that practitioners building AI-enabled systems regularly struggle to translate high-level principles into concrete engineering decisions [12, 25]. Existing tool support does not help much: enterprise compliance platforms carry significant overhead, and general-purpose LLM assistants will generate requirement text on demand, but without grounding in authoritative regulatory sources there is little basis for trusting what they produce [15]. Securing industry participation in *HealthAISys* research is itself a significant challenge: regulatory sensitivity, confidentiality obligations, and organisational risk aversion mean it is very difficult to get access to practitioners working on real projects.

To understand this gap from a practitioner perspective, we previously conducted a semi-structured interview study with 17 *HealthAISys* stakeholders and practitioners spanning clinicians as end users and software engineers as developers (currently under review at another venue). Participants consistently reported difficulty translating regulatory provisions into implementable requirements, uncertainty about which ethical dimensions applied to their specific product context, and a lack of lightweight tooling that could fit into their existing workflows without imposing compliance-specialist overhead. These findings directly shaped the design decisions of the tool described in this paper.

To address this, we developed *HealthAI Ethics Assistant* grounded in a knowledge base rather than open-ended LLM generation. The knowledge base combines regulatory guidance from the EU AI Act and the NIST AI RMF with practitioner-identified ethical concerns drawn from the interview study described above and scientific literature [13]. This grounds the tool in explicit, traceable sources: every generated requirement can be linked to a specific regulatory provision or a real practitioner concern, rather than general pre-training data. The tool is intentionally scoped as a working prototype, and this paper reports an exploratory evaluation with a practitioner – an R&D engineer at a medical device manufacturer who is also a co-author of this paper. The evaluation was designed to test feasibility and refine the study protocol ahead of a multi-site study. We hence make no claims of generalisability.

This paper describes the design, implementation, and exploratory practitioner evaluation of the *HealthAI Ethics Assistant*, an AI-enabled RE tool built to support healthcare AI software teams. The tool guides practitioners through a Create–Validate–Compare workflow and grounds every output in a structured knowledge base.

Contributions. This paper makes the following contributions:

- (1) A description of the design and implementation of the HealthAI Ethics Assistant, including its Create–Validate–Compare interaction model and traceable regulatory knowledge base informed by practitioner findings from our interview study.

- (2) Preliminary evaluation findings from an exploratory practitioner case study with an R&D engineer at a medical device manufacturer, collected using the Technology Acceptance Model [9] and the System Usability Scale [8].
- (3) Lessons for designing AI-enabled RE tools in regulated healthcare domains.
- (4) We make the tool demo interface [1] and the source code repository [2] publicly available.

2 Background and Related Work

AI ethics guidance has expanded rapidly, with broad frameworks such as the OECD AI Principles [19] and UNESCO Recommendation [24] establishing shared values around transparency, fairness, and accountability. Jobin et al. found that most guidelines offer little concrete guidance on how to act on them [14], and Mittelstadt stated that without implementation support, principle-level guidance risks becoming ethics washing rather than meaningful change [17]. More recently, governance has shifted to enforceable regulation: the NIST AI Risk Management Framework [22] provides a risk-based structure across Govern, Map, Measure, and Manage functions, while the EU AI Act establishes the first legally binding regional AI regulation, classifying healthcare AI as high-risk. Translating these frameworks into actionable engineering work remains a recurring challenge for practitioners [5, 16].

Within RE, Vogelsang and Borg argued that traditional RE methods are poorly suited to AI systems where quality attributes such as fairness and explainability are difficult to specify in verifiable terms [25]. Habibullah et al. surveyed practitioners on non-functional requirements for ML and found that fairness, transparency, and safety are frequently identified as important but rarely specified concretely [12]. Nalchigar et al. reported similar difficulties in a healthcare ML modelling case study [18]. LLM-assisted RE has emerged as a possible solution: a recent systematic literature review found 136% growth in LLM-for-RE studies between 2023 and 2024 [26], with tools like ReqBrain showing fine-tuned LLMs can generate practitioner-relevant requirements [11]. However, a trust problem remains: LLM outputs are difficult to audit without grounding in authoritative sources, motivating retrieval-augmented approaches [15]. This is the gap our paper aims to address.

Tool support for regulatory compliance is well established in other regulated domains. In privacy and data protection, a line of tools automates compliance checking against the GDPR: Torre et al. use conceptual models to check system compliance [23], Amaral et al. check the completeness of privacy policies against regulatory provisions [4], and Sleimi et al. [21] and later Abualhaija et al. [3] use semantic legal metadata to support downstream analysis, building on earlier results from Breaux and Anton on systematically extracting rights and obligations from regulatory text [7]. Two properties separate this body of work from our setting. It is predominantly verification-oriented, which means that an artefact already exists and the tool checks it against a source. And it targets provisions specific enough to formalise, whereas our sources span binding regulation and principle-level guidance that resists it.

Positioning relative to prior work. Unlike prior LLM-for-RE tools that evaluate outputs in isolation, our tool is grounded in a regulatory knowledge base and practitioner findings from our interview study, enabling traceability from the generated requirement to the source. This positions our tool as an auditable compliance aid rather than a generative assistant.

3 Methodology

This section describes the design rationale and technical implementation of the HealthAI Ethics Assistant.

3.1 Motivating Interview Study

Before designing the tool, we conducted a semi-structured interview study with 17 healthcare AI practitioners: five healthcare professionals, six healthcare researchers, and six software engineers, recruited from IT companies, clinics, hospitals, and universities in China (8), Australia (8), and India (1), with 1 to 24 years of professional experience. Interviews were conducted between February and July 2025, recorded, transcribed, and analysed using thematic synthesis. The study built on our previous systematic mapping study [13], which identified five ethical dimensions. Participants ranked these by priority, yielding privacy (mean 1.2), safety (2.2), bias (3.3), accountability (3.7), and transparency (4.5), and all five recurred in the interview data as sources of practical difficulty. These findings are encoded in `findings.json` and injected into the tool's Compare workflow to ground requirement refinement in real practitioner experience rather than synthetic examples. The full interview study is currently under review at another venue; this paper focuses on the tool that the study motivated and its initial evaluation.

3.2 Design Rationale

Our design decision was to ground the assistant's structured workflow in explicit, curated knowledge sources. This was motivated by a practical concern: in compliance-oriented requirements work, practitioners need to know *where* a requirement comes from, not just what it says. The knowledge base was assembled from two sources. The first is a set of regulatory frameworks, specifically the EU AI Act [10] and the NIST AI Risk Management Framework [22], they are encoded as structured JSON covering article titles, summaries, implementation guidance, applicable ethical dimensions, and source links. The second is the set of practitioner-identified concerns drawn from the interview study described in Section 3.1, stored as `findings.json` and covering the five ethical dimensions of privacy, safety, bias, transparency, and accountability. These five dimensions were adopted from an earlier systematic mapping study of generative AI ethics [13], and their recurrence in the interview data provided independent empirical validation that they are the right organising structure for the tool's outputs.

The two knowledge sources are used at different stages of the workflow through a lightweight structured RAG pipeline rather than embedding-based vector retrieval. For Create and Validate interactions, the backend retrieves regulatory entries using structured metadata such as ethical dimension and framework identifier; for Compare interactions, it retrieves practitioner findings by stakeholder role and ethical dimension. Retrieved records are formatted as grounding context and injected into a mode-specific prompt with explicit instructions that generated requirements and validation checks cite only the supplied guideline identifiers. The model returns structured JSON, which the backend post-processes to enrich citation metadata before rendering reviewable requirement cards.

We selected sources that are authoritative rather than advisory, that were formally published when we built the knowledge base, and that cover the regulations our interview participants raised most often. The knowledge base is a fixed snapshot: entries are edited by hand, nothing records which version of a framework an entry came from, and the tool does

not check whether two sources disagree. When two frameworks cover the same obligation but set different standards, both entries are injected and the generated requirement cites both, leaving the practitioner to decide which applies. This works at the current scale, but covering more jurisdictions would first require version tracking and automatic conflict detection.

3.3 Interaction Model

The tool enforces practitioner interaction through a **Create–Validate–Compare** workflow. In **Create** mode, the tool generates ethical concerns and implementable software requirements in response to a project or scenario description. In **Validate** mode, it checks an existing requirement against the built-in knowledge base and flags gaps or misalignments with regulatory guidelines. In **Compare** mode, it takes a draft requirement, compares it against practitioner findings from the interview study, and proposes a refined version. The backend automatically classifies user messages into one of these three modes via the `POST /ethics/chat` endpoint, removing the need for practitioners to explicitly select a mode.

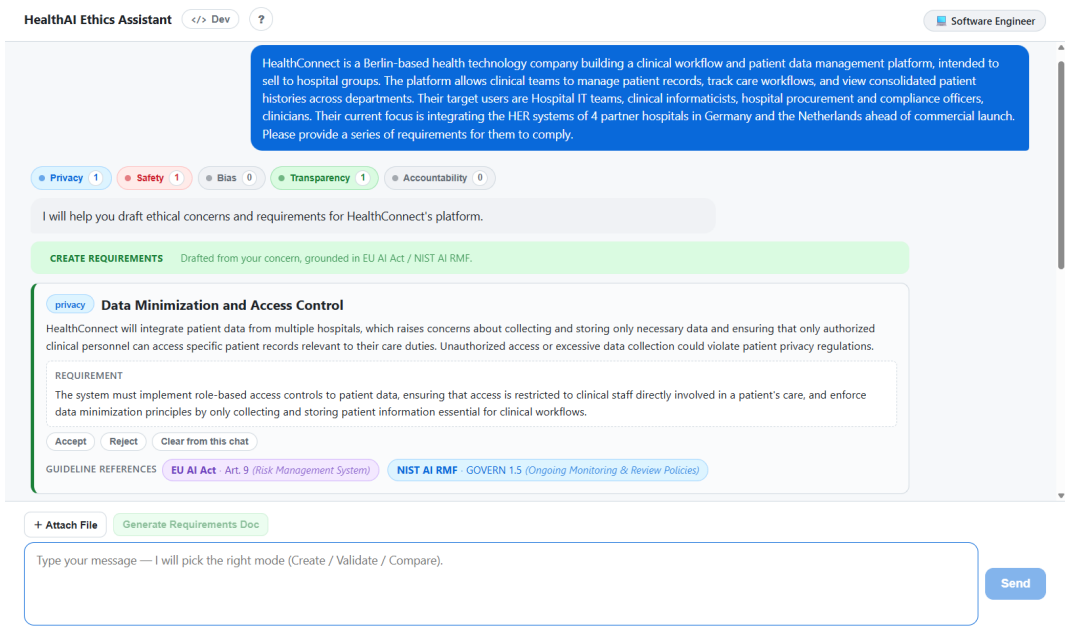
To start a session, the practitioner clicks *New Chat*, which opens a configuration modal. Here, the practitioner selects a topic, assigns a chat name, chooses a persona, and optionally provides a project context. Three fixed personas are available: Healthcare Professional, Software Engineer, and Healthcare Researcher with a *Let System Decide* option that presents a 20-question behavioural questionnaire to help the user determine their persona. The frontend then passes a compact JSON persona description to the backend as prompt context, allowing the LLM to adapt the tone, specificity, and framing of generated requirements to the practitioner’s background.

Each generated requirement is presented as a structured card that displays the ethical dimension covered, the requirement statement, and clickable regulation-reference blocks linking directly to specific EU AI Act articles and NIST AI RMF controls. Users can mark each requirement as Accept, Reject, or Clear, with accepted requirements exportable as a structured `.docx` document grouped by ethical dimension.

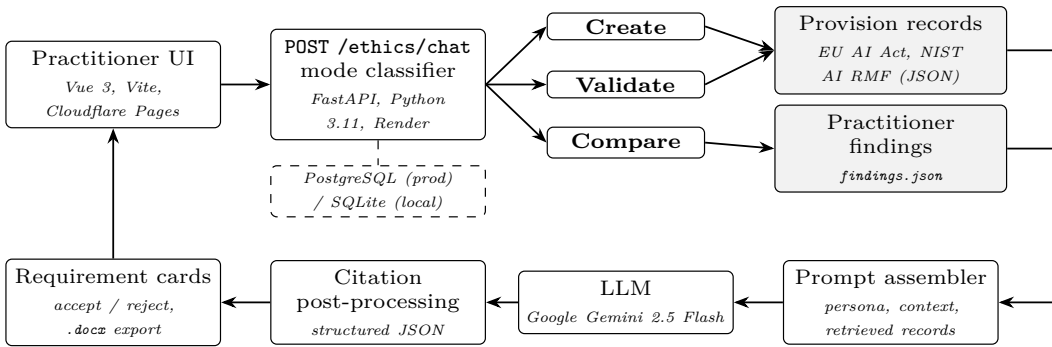
Figure 2 summarises the technology workflow. The backend is built with FastAPI and exposes a REST API consumed by a Vue 3 single-page application hosted on Cloudflare Pages. The LLM layer uses Google Gemini (`gemini-2.5-flash` by default), invoked via the `ethics_service.py` module, which constructs mode-specific grounded prompts by combining the selected persona context, project context, relevant knowledge-base entries, requirements preferences, and the current user message. Session data, chat messages, and requirement preferences are persisted in PostgreSQL in production and SQLite for local development. A lightweight document grounding mechanism available in developer mode allows practitioners to upload reference documents (`.md`, `.txt`, or `.json`) that are prepended to the LLM prompt as additional context. The tool can be accessed via our demo site [1]. The tool source code, knowledge base, and example session data are available in a replication package [2].

4 Exploratory Practitioner Study

To assess the HealthAI Ethics Assistant in a realistic industrial setting, we conducted an exploratory practitioner study with an R&D engineer at BMC Medical Co., Ltd., a Chinese medical device manufacturer specialising in respiratory healthcare. BMC has operated for over 20 years, ranks second globally in CPAP device sales, holds FDA certification, and distributes across more than 140 countries and regions [6]. The expert develops respiratory medical devices including ventilators and CPAP systems, products that fall within the EU AI Act high-risk category when AI-enabled. The expert (Cheng Zhang, co-author) had two



■ **Figure 1** A sample session showing a generated requirement card with traceable references.



■ **Figure 2** Architecture and Create–Validate–Compare workflow of the HealthAI Ethics Assistant.

distinct roles in this research programme. In the *design phase*, he contributed to industrial problem framing by reviewing tool requirements derived from the interview study, validating that the Create–Validate–Compare workflow reflected realistic compliance needs in a medical device context, and advising on which regulatory dimensions were most pressing for his organisation. In the *evaluation phase*, he participated as the first practitioner evaluator, conducting an unstructured tool-use session followed by structured post-use assessment. This dual role reflects the industry–academia collaboration model for our project. We acknowledge that his involvement in the design phase is a source of response bias. To limit it, he was not shown the tool’s implementation, the contents of its knowledge base, or any generated output before the session, and the session used a scenario he selected himself rather than one prepared by the research team. These measures reduce bias but cannot eliminate it; therefore, we treat the ratings as indicative rather than as an independent assessment.

The case study combined a tool-use session and a post-use evaluation using the Technology Acceptance Model (TAM) [9] and the System Usability Scale (SUS) [8], complemented by open-ended qualitative feedback. Qualitative feedback was collected through four open-ended

prompts administered after the structured instruments, asking which aspects of the tool were most and least valuable, what would prevent its adoption in the evaluator's organisation, and what functionality was missing. Responses were recorded in writing, and the first author grouped them into the observations. This approach is consistent with established practice in empirical software engineering for early-stage prototype evaluation [20]. Prior to the session, the expert received a brief overview of the tool's purpose and was asked to explore it freely using a scenario from their own engineering work. The expert chose to describe an in-development dual-level non-invasive home ventilator. The session continued over three interactions, each automatically classified by the backend into a different workflow mode.

Create (initial scenario). The expert described the target product with details: pressure range 4–30 cmH₂O, humidification 10 g/L, supporting CPAP, auto-CPAP, S, auto-S, PCV, and ASV modes, with full respiratory-event logging. The tool returned a Safety requirement on reliable operation in home environments, grounded in EU AI Act Art. 9 (Risk Management System) and Art. 15 (Accuracy, Robustness and Cybersecurity), and NIST AI RMF MEASURE 2.3 (AI System Performance and Validity). A second requirement addressed Privacy, focused on secure handling of respiratory event data with encryption at rest and in transit and anonymisation for research use. Both requirements were accepted by the expert.

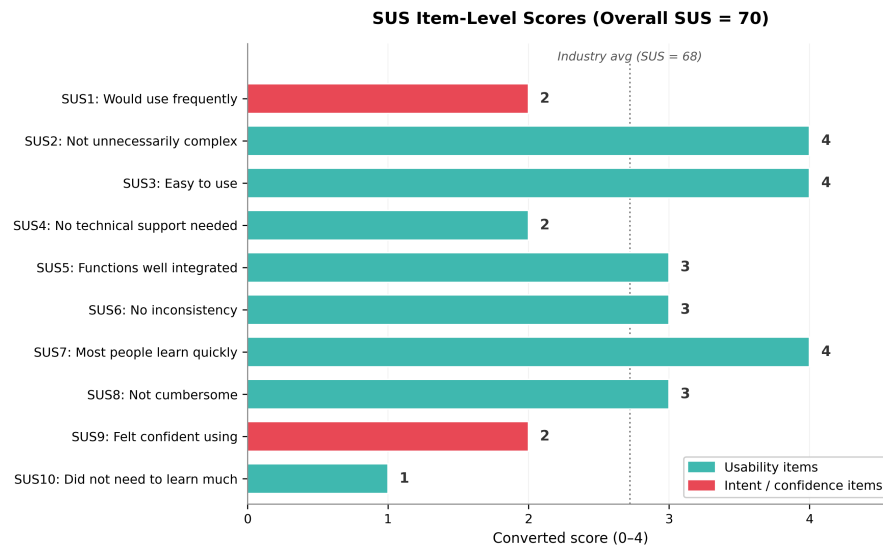
Create (extended context). The expert added that the product is developed in China but distributed across China, the US, India, and Europe, with EMC IEC 61000 testing and CE certification in place. The tool returned a Bias requirement on algorithmic performance across demographic, genetic, and cultural populations (EU AI Act Art. 10; NIST AI RMF MEASURE 2.11) and a Transparency requirement on region-specific regulatory disclosure (Art. 13; MAP 1.5).

Compare. The expert further asked: *“I want to know what detailed additional certifications are needed in the US, Europe, and India.”* The backend classified this as a Compare interaction. The tool found three practitioner findings from the interview study knowledge base on the accountability dimension: a perspective from a healthcare professional discussing AI malpractice liability across the supply chain, an interview quote from a clinician on human oversight requirements, and a view from a healthcare researcher on model versioning and reproducibility. Each finding was annotated with a “Consider:” prompt suggesting concrete refinements to the certification scope.

Evaluation Results

Given that this is a preliminary evaluation with a single expert, all results are treated as exploratory rather than conclusive. We organise findings across three lenses: system usability (SUS), technology acceptance (TAM), and open qualitative feedback.

For the System Usability Scale (SUS), the composite score was 70. With a single respondent, this figure carries no inferential weight, and we make no comparison against population benchmarks, which are means across many respondents and studies. The item-level pattern in Figure 3 is more informative. Six of the eight usability items were rated at 3 or 4: the evaluator found the tool not unnecessarily complex (SUS2=4), easy to use (SUS3=4), quick for most people to learn (SUS7=4), well integrated (SUS5=3), consistent (SUS6=3), and not cumbersome (SUS8=3). Of the remaining two, he was neutral on whether technical support would be needed (SUS4=2) and disagreed that the tool could be used without first learning a good deal (SUS10=1), the lowest of the SUS items. Both



■ **Figure 3** SUS item-level scores.

concern onboarding rather than the interface, and both are consistent with the persona configuration difficulty reported under TAM Q3 below. Across all eight usability items the average was 3.0 out of 4. The two intent and confidence items sat at the scale midpoint (SUS1 = 2, SUS9 = 2); the evaluator attributed both explicitly to role mismatch rather than to dissatisfaction with the tool, noting that compliance requirement drafting falls outside the primary responsibilities of an R&D engineer.

For the Technology Acceptance Model (TAM), the most prominent TAM finding was Q10: traceable regulatory references received a maximum score of 5, the highest rating in the evaluation. The expert identified this as the most valuable aspect of the tool in open feedback: *“Traceable regulatory specifications, indexing of related legal clauses and international standards make me understand clearly which legal provisions are involved.”* This directly validates the design decision to ground the tool in explicit knowledge sources rather than open-ended LLM outputs. Ease-of-use items rated positively (Q2 = 4 on learning without training; Q1 = 3 on navigation; Q4 = 3 on the accept/reject interface). Q3 (clarity of persona onboarding) received a 2, with the expert noting that the three role categories are *“a bit vague, better to define product engineer, software engineer, hardware engineer, and market manager separately.”* Perceived Usefulness items (Q5–Q8) were rated neutral at 3 each. Q11 (transparency of reasoning) received a 3, suggesting that while regulatory grounding was well received, the explanation of *why* a specific requirement was generated could be made more explicit. The two Intention to Use items were the lowest TAM ratings (Q12 = 2 on future use; Q13 = 2 on recommending the tool to colleagues), again reflecting role fit; the expert noted these would likely score higher for a product engineering or regulatory affairs respondent. Table 1 summarises the full TAM responses. Two signals in Table 1 are actionable. Eight of the thirteen items sat at the midpoint, which the evaluator attributes to requirements that restate knowledge already standard for a device engineer and to the absence of device-level parameters. And the gap between Q10 = 5 and Q11 = 3 suggests source citation and reasoning explanation are separate properties: the tool shows which provision a requirement links to, but not why it was selected, which will be the priority for the next iteration.

■ **Table 1** TAM evaluation responses (1 = Strongly Disagree, 5 = Strongly Agree).

Aspect	Item	Question (abbreviated)	Score
PEOU	Q1	Easy to use and navigate	3
	Q2	Easy to learn without training	4
	Q3	Persona onboarding was clear	2
	Q4	Accept/reject interface was intuitive	3
PU	Q5	Improves compliance drafting	3
	Q6	Relevant and useful information	3
	Q7	Five-dimensional indicator useful	3
	Q8	Mixed persona produced useful output	3
TT	Q9	Trusted the generated requirements	3
	Q10	Traceable references increased trust	5
	Q11	Transparent in why output generated	3
ITU	Q12	Would use in future projects	2
	Q13	Would recommend to colleagues	2

Open Qualitative Feedback. The expert identified traceable regulatory specifications and the indexing of relevant legal clauses as the most valued aspect of the tool, stating that it enables a clear understanding of the legal obligations involved in healthcare AI development. This directly aligns with the strong trust rating observed for Q10.

The expert identified three barriers to adoption. First, the tool’s built-in knowledge base covers EU and US regulatory frameworks, but R&D teams operating under other jurisdictions require broader coverage. The tool’s document-grounding mechanism allows practitioners to upload additional regulatory documents, but it places the burden of sourcing and formatting them on the practitioner rather than providing built-in coverage. Second, certain generated requirements constitute standard industry knowledge for device engineers and offer no practical benefit to that role. Third, device-level parameters such as pressure ranges for CPAP and ventilator modes are essential for generating requirements under real product constraints, and their absence limits the tool’s applicability for R&D work.

5 Discussion

The evaluation shows that the tool functions well as an interface, but its practical value depends heavily on the practitioner’s role. The expert’s work as an R&D engineer focuses on device-level details such as pressure tolerances, hardware constraints, and component specifications. The HealthAI Ethics Assistant is designed to generate compliance requirements based on regulatory frameworks, which explains why the expert found the tool conceptually useful but limited in direct applicability. In medical device organisations, compliance work typically falls to product engineers and regulatory affairs teams; the tool would likely be perceived as more useful in those hands. The feature that worked consistently across the evaluation was traceable regulation references: knowing that a generated requirement links back to a specific guideline makes it far easier to defend in a design review or audit. The five ethical dimensions are not equally easy to turn into requirements. Privacy, safety, and bias correspond to provisions that state concrete engineering obligations, and the tool handled them readily. Accountability is harder because obligations regarding liability and oversight depend on who within the organisation is answerable, and until that is decided, there is little for the tool to specify. Our interview study showed the same pattern, with accountability ranked fourth of five, and in the evaluation session it appeared only through Compare, as retrieved findings rather than generated requirements.

5.1 Design Recommendations

R1: Expand the built-in regulatory coverage. The current knowledge base covers the EU AI Act and NIST AI RMF, but does not reflect the wider regulatory landscape that international teams must navigate. Future versions should incorporate Chinese AI governance standards, relevant IEC/ISO medical device standards, and region-specific disclosure requirements. A curation roadmap driven by practitioner-reported gaps would help prioritise which frameworks to add first. The document upload feature is a viable short-term workaround, but should be supplemented with guided templates to reduce the formatting burden on practitioners.

R2: Refine role and context awareness. The persona questionnaire treats SE as a single category, but compliance responsibility varies significantly across roles. Future versions should offer a finer-grained role taxonomy (e.g., R&D engineer, product engineer, clinical lead) and adapt the tool's positioning accordingly: for roles with direct compliance responsibility, the tool should function as a primary drafting aid; for roles with peripheral compliance involvement, it should present itself as a reference and awareness resource rather than a workflow tool. This shift in framing, rather than a change in functionality would likely improve perceived usefulness for roles like the expert's without requiring new features.

5.2 Lessons Learned

L1: Traceability builds trust faster than reasoning quality . The evaluation showed that what mattered most to the practitioner was not the quality of the LLM's reasoning or the breadth of its output, but that every generated requirement could be traced to a specific regulatory article or practitioner finding. This trust was achieved through structured JSON prompt grounding, a lightweight mechanism that does not require a full RAG pipeline. The practical implication is that teams building compliance-oriented LLM tools should treat source traceability as a first-class design concern, not an audit feature added later.

L2: Role fit determines adoption, not usability. Another finding from the evaluation was not an interface problem but a positioning problem: the tool was used by the right *organisation* but the wrong *role*. This is predictable for compliance tools in large organisations, where the person with domain knowledge (R&D engineer) and the person with compliance responsibility (product or regulatory engineer) are different people. For teams building AI-enabled RE tools in regulated domains, the design question “who will use this?” should be answered at the level of job role and compliance responsibility, not just domain.

L3: Practitioners stress-test scope before evaluating features. The expert didn't explore the tool's features systematically. He threw a real, hard problem at it and watched where it broke. This is how practitioners evaluate tools. The implication is that graceful boundary behaviour matters as much as core functionality: a tool that clearly signals its limits is trusted more than one that produces plausible-sounding output beyond its competence.

L4: Embedded ethics knowledge surfaces blind spots. The expert flagged some requirements as “standard industry knowledge”. However, what is standard to a 20-year veteran might not be to a new hire, a contractor, or developer teams in healthcare from adjacent domains. Ethics and compliance requirements that feel obvious to experts are genuinely non-obvious to others. AI-enabled tools that encode ethical dimensions may matter less as expert aids and more as organisational awareness tools – surfacing ethical obligations to people who would not otherwise know to ask about them.

6 Threats to Validity

This study has several limitations that should be considered when interpreting the findings. First, the evaluation involved a single expert, which limits the generalisability of the results; findings are exploratory rather than validated conclusions. The single-expert design was a deliberate choice for this phase: the goal was to assess fit between the tool and real industrial practice before investing in a broader multi-site study, consistent with early-stage prototype evaluation guidelines [20]. Furthermore, as noted in the introduction, securing any industry participation in healthcare AI research carries significant access barriers, making even a single engaged practitioner from a real medical device organisation a meaningful empirical contribution at this stage. Second, the expert was an R&D engineer, whereas the tool is primarily designed for product engineers and regulatory affairs specialists. Our future study will address these three limitations. We will recruit practitioners from several healthcare AI organisations, selected by compliance responsibility rather than technical domain, so the sample includes product engineers, regulatory affairs specialists, and clinical leads alongside R&D engineers, none of whom will have helped design the tool. Each will complete two sessions, one on a shared scenario and one on a product from their own work. Third, the evaluator was also briefly involved in the tool design phase, which may have introduced response bias toward positive assessments. Fourth, the knowledge base currently covers only the EU AI Act and NIST AI RMF, which may not reflect the regulatory context of teams operating under other jurisdictions. Generalisability is limited in three respects: one region, one role, and one product, the last being a single home ventilator chosen by the evaluator. These results may not hold in other regulatory contexts, for other roles, or for other kinds of product. Future iterations will incorporate Chinese AI governance standards and relevant IEC/ISO medical device standards.

7 Conclusion

This paper presents the design, implementation, and initial practitioner evaluation of the HealthAI Ethics Assistant, an AI-enabled requirements engineering tool built to help healthcare AI software teams translate regulatory guidance into concrete ethical requirements. The tool was motivated by a semi-structured interview study with healthcare SE practitioners spanning clinicians and software engineers, whose reported difficulties in mapping regulatory obligations to engineering tasks directly shaped the knowledge base design and workflow structure (full study currently under review). The tool is grounded in a structured knowledge base combining the EU AI Act, the NIST AI RMF, and practitioner findings from that study, and supports a Create–Validate–Compare workflow with traceable regulatory references.

An initial evaluation was conducted with an R&D engineer from BMC Medical Co., Ltd., using the Technology Acceptance Model and the System Usability Scale, along with open-ended qualitative feedback. The tool received generally positive usability ratings and a strong endorsement of its traceable regulatory references, which were identified as its most trustworthy feature. The main adoption challenge was not basic usability but role fit. From this, the paper derived practical design recommendations on expanding regulatory and ethics coverage and improving role awareness, alongside transferable lessons on the importance of source traceability and clear workflow positioning for AI-enabled tools in healthcare.

Future work will extend the evaluation to a multi-case study covering multiple roles across several healthcare organisations and broaden the knowledge base to include additional regulatory frameworks. Together, these findings suggest that AI-enabled assistants can

meaningfully support the translation of ethical AI principles into practical requirements engineering work, provided they are designed as auditable, human-in-the-loop tools that align with the practitioners' actual responsibilities.

References

- 1 HealthAI ethics assistant demo. <https://healthai-ethics.jche0515.workers.dev/>.
- 2 HealthAI ethics assistant replication package. <https://anonymous.4open.science/r/healthai-ethics-assistant-artifact-84B1/>.
- 3 Sallam Abualhaija, Chetan Arora, Amin Sleimi, and Lionel C Briand. Automated question answering for improved understanding of compliance requirements: A multi-document study. In *2022 IEEE 30th international requirements engineering conference (RE)*, pages 39–50. IEEE, 2022. doi:10.1109/RE54965.2022.00011.
- 4 Orlando Amaral, Sallam Abualhaija, Damiano Torre, Mehrdad Sabetzadeh, and Lionel C Briand. Ai-enabled automation for completeness checking of privacy policies. *IEEE Transactions on Software Engineering*, 48(11):4647–4674, 2021. doi:10.1109/TSE.2021.3124332.
- 5 Laura Arbelaez Ossa, Giorgia Lorenzini, Stephen R Milford, David Shaw, Bernice S Elger, and Michael Rost. Integrating ethics in ai development: a qualitative study. *BMC Medical Ethics*, 25(1):10, 2024.
- 6 BMC Medical Co., Ltd. About BMC medical, 2025. URL: <https://en.bmc-medical.com/en/about-bmc/index>.
- 7 Travis Breaux and Annie Antón. Analyzing regulatory rules for privacy and security requirements. *IEEE transactions on software engineering*, 34(1):5–20, 2008. doi:10.1109/TSE.2007.70746.
- 8 John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.
- 9 Fred D Davis, Richard P Bagozzi, and Paul R Warshaw. Technology acceptance model. *J Manag Sci*, 35(8):982–1003, 1989.
- 10 European Parliament and Council of the European Union. Regulation (EU) 2024/1689 (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689, 2024. URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.
- 11 Mohammad Kasra Habib, Daniel Graziotin, and Stefan Wagner. Reqbrain: Task-specific instruction tuning of llms for ai-assisted requirements generation. *arXiv preprint arXiv:2505.17632*, 2025. doi:10.48550/arXiv.2505.17632.
- 12 Khan Mohammad Habibullah, Gregory Gay, and Jennifer Horkoff. Non-functional requirements for machine learning: Understanding current use and challenges among practitioners. *Requirements Engineering*, 28(2):283–316, 2023. doi:10.1007/S00766-022-00395-3.
- 13 Yutan Huang, Chetan Arora, Wen Cheng Huong, Tanjila Kanij, Anuradha Madugalla, and John Grundy. Ethical concerns of generative ai and mitigation strategies: A systematic mapping study. *Applied Soft Computing*, page 114789, 2026. doi:10.1016/J.ASOC.2026.114789.
- 14 Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399, 2019. doi:10.1038/S42256-019-0088-2.
- 15 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
- 16 Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, David Douglas, and Conrad Sanderson. Software engineering for responsible ai: An empirical study and operationalised patterns. In *Proceedings of the 44th International Conference on Software Engineering: Software Engineering in Practice*, pages 241–242, 2022. doi:10.1109/ICSE-SEIP55303.2022.9793864.
- 17 Brent Mittelstadt. Principles alone cannot guarantee ethical ai. *Nature machine intelligence*, 1(11):501–507, 2019. doi:10.1038/S42256-019-0114-4.

- 18 Soroosh Nalchigar, Eric Yu, and Karim Keshavjee. Modeling machine learning requirements from three perspectives: a case report from the healthcare domain. *Requirements Engineering*, 26(2):237–254, 2021. doi:10.1007/S00766-020-00343-Z.
- 19 OECD. OECD AI principles overview, 2024. URL: <https://oecd.ai/en/ai-principles>.
- 20 Per Runeson and Martin Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, 14(2):131–164, 2009. doi:10.1007/S10664-008-9102-8.
- 21 Amin Sleimi, Nicolas Sannier, Mehrdad Sabetzadeh, Lionel Briand, and John Dann. Automated extraction of semantic legal metadata using natural language processing. In *2018 IEEE 26th International Requirements Engineering Conference (RE)*, pages 124–135. IEEE, 2018. doi:10.1109/RE.2018.00022.
- 22 Elham Tabassi. Artificial intelligence risk management framework (ai rmf 1.0), 2023.
- 23 Damiano Torre, Ghanem Soltana, Mehrdad Sabetzadeh, Lionel C Briand, Yuri Auffinger, and Peter Goes. Using models to enable compliance checking against the gdpr: an experience report. In *2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems (MODELS)*, pages 1–11. IEEE, 2019. doi:10.1109/MODELS.2019.00-20.
- 24 UNESCO. Recommendation on the ethics of artificial intelligence, 2021. URL: <https://unesdoc.unesco.org/ark:/48223/pf0000381137>.
- 25 Andreas Vogelsang and Markus Borg. Requirements engineering for machine learning: Perspectives from data scientists. In *2019 IEEE 27th international requirements engineering conference workshops (REW)*, pages 245–251. IEEE, 2019. doi:10.1109/REW.2019.00050.
- 26 Mohammad Amin Zadenoori, Jacek Dąbrowski, Waad Alhoshan, Liping Zhao, and Alessio Ferrari. Large language models (llms) for requirements engineering (re): A systematic literature review. *arXiv preprint arXiv:2509.11446*, 2025. doi:10.48550/arXiv.2509.11446.