AI-Assisted Secondary Use of Clinical Research Data: A Three-Dimensional Personalization Framework for Scientific Publication Drafting

Elisavet Persefoni Kanidou
Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
elisavetkanidou@gmail.com

Nikolaos Nikolaidis
Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
nnikolaidis@uom.edu.gr

Konstantina Tsimpita
Aristotle University of Thessaloniki
Thessaloniki, Greece
constantina.tsimpita@gmail.com

Apostolos Ampatzoglou
Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
apostolos.ampatzoglou@gmail.com

Alexandros Chatzigeorgiou
Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
achat@uom.edu.gr

Panagiotis Bamidis
School of Medicine
Aristotle University of Thessaloniki
Thessaloniki, Greece
bamidis@med.auth.gr

Evdokimos Konstantinidis
School of Medicine
Aristotle University of Thessaloniki
Thessaloniki, Greece
evdokimosk@gmail.com

***Abstract*—The secondary use of clinical research data for scientific publication remains a significant bottleneck in medical research, as collected datasets are rarely transformed into disseminated findings due to the complexity of the data and the writing process. In this paper, we propose an open-source three-dimensional personalization framework (Lab, Personal, Global), to dynamically control these knowledge influences in the paper preparation. The framework extracts ML-based writing style features from reference papers and constructs adaptive prompts based on the set influence levels. The system supports four AI vendors (Groq Llama 3.3 70B, Google Gemini, OpenAI GPT-4, and local GPT-OSS 120B) through a unified abstraction layer. An evaluation with 20 researchers showed significant improvements in writing consistency over baseline methods, along with strong usability scores (SUS 80.78). The framework was validated using real-world clinical data from the RAISE platform, demonstrating its potential to accelerate the transformation of research datasets into publishable scientific output.**

***Index Terms*—medical writing, AI personalization, secondary data use, natural language processing, writing style analysis, health data lifecycle**

I. Introduction

Clinical research generates vast amounts of data every day, yet a significant proportion of these datasets never reach publication. Empirical studies estimate that approximately half of completed clinical trials remain unpublished [1], representing a major loss of scientific value and a threat to evidence-based medicine. A key barrier is the complexity of transforming collected data into a well-structured, field-compliant manuscript, a process that demands alignment with institutional conventions, personal writing experience, and current field standards simultaneously. The RAISE platform, a European funded research infrastructure, exemplifies this challenge: researchers collect and manage high-quality data (including clinical data) but lack AI-assisted tools to support the subsequent publication process.

Writing a medical research paper is not a straightforward task, especially when the author has to respect the habits of their lab, keep a consistent personal style, and at the same time follow whatever the field considers best practice at the moment. The challenge is that no existing AI system can meet all three requirements simultaneously. Grammarly[1], for instance, can correct grammar and enhance style, but it lacks awareness of the specialized conventions used in a university hospital's cardiology lab. GPT-4 [2] can produce high-quality prose in zero-shot and few-shot settings, but maintaining consistent stylistic preferences across independent interactions typically requires re-specifying user instructions. To our knowledge, no existing mainstream AI writing system provides explicit, user-controlled multi-source influence parameters (e.g., adjustable lab, personal, and global weighting).

Given the aforementioned gap, two challenges in AI-assisted academic writing are identified. The first is *multi-dimensional personalization*: designing an AI assistant that dynamically balances three knowledge sources while remaining transparent and controllable. This requires: (a) quantitative extraction of

[1]https://www.grammarly.com

writing style features from heterogeneous paper collections; (b) interpretable parameterization for adjusting influence levels without ML expertise; (c) efficient context aggregation within LLM token limits (4K–8K tokens); and (d) real-time adaptation as users upload new references or adjust preferences. Architecturally, the solution demands vendor-agnostic design supporting cloud providers and on-premises deployment, real-time collaboration, and role-based access control.

The second challenge, which is complementary to the first, has to do with *dataset repurposing and research opportunity identification*. It is very common for a research team to finish a clinical study and then never use the collected data again, even though those data could answer other questions, for example, a subgroup analysis or a look into adverse event patterns. The issue is that figuring out what else can be done with an existing dataset requires someone who knows both the data and the current literature well enough. None of the existing AI tools are built to do this.

To address both challenges, the contributions of this paper are the following. First, a three-dimensional personalization model is introduced with interpretable sliders (Lab, Personal, Global; 0–10) that dynamically construct AI prompts via guidance functions. Second, a multi-provider AI architecture is implemented, supporting Groq (Llama 3.3 70B), Google Gemini 2.5 Flash, OpenAI GPT-4, and local GPT-OSS 120B through a unified abstraction layer. Third, the framework can be used for data-related tasks: researchers used it to check if their datasets had gaps, to spot inconsistencies, and to verify whether they were meeting the quality standards expected in their domain. This matters significantly in the context of responsible data lifecycle management, where ensuring data quality, enabling secondary use, and supporting FAIR (Findable, Accessible, Interoperable, Reusable) [3] data principles are central priorities. The proposed framework directly addresses this gap by integrating data quality verification and research opportunity identification into the AI-assisted writing workflow [4]. Fourth, an empirical evaluation is conducted, with a SUS score of 80.78 with 20 participants, and in domain-specific study for real-world data collected through the RAISE platform, examining clinical terminology consistency demonstrated an improvement in writing features over baseline.

The rest of the paper is organized as follows. Section II presents the related work. In Section III, the proposed system design is described, including the three-dimensional personalization model, the NLP feature extraction pipeline, and the multi-provider AI architecture. Section IV details the evaluation of the framework, which is carried out in two stages: a system usability analysis and a medical writing experiment. Section V then addresses the implications, limitations, and ethical aspects. Lastly, Section VI provides the conclusion of the paper.

II. Related Work

In this section, existing studies and background information for this research effort are presented. First, studies related to AI writing assistants are discussed; second, personalization approaches in NLP, third, medical research platforms; and finally, writing style analysis methods.

A. AI Writing Assistants

Rule-based tools like Grammarly and ProWritingAid[2] generate identical suggestions regardless of domain, making them unsuitable for specialized medical writing. LLM-based systems such as Claude [5] and ChatGPT [2] offer better contextual awareness but do not persist user-specific settings across sessions. Alignment-based approaches such as InstructGPT [6] improve controllability but do not provide explicit multi-source personalization. Overleaf[3] and Google Docs[4] support collaborative editing but provide no AI customization for medical writing workflows.

B. Personalized NLP Systems

Existing personalization approaches address only single dimensions: StyleTransfer [7] adjusts formality but ignores domain context; persona-based dialogue systems [8] learn from interaction history but offer no explicit user control, few-shot learning [9] replicates style from examples but requires re-execution each session. Parameter-efficient fine-tuning approaches such as LoRA [10] enable model adaptation but require retraining and therefore cannot provide real-time, user-controlled personalization. None support simultaneous multi-source influence control or provide quantitative feedback on style alignment.

C. Medical Research Platforms

Reference managers (ResearchGate[5] and Mendeley[6]) and collaborative editors (Authorea[7]) lack AI writing support. Clinical data platforms (REDCap [11], OpenClinica[8]) support data collection but not dissemination. No existing platform addresses the full data lifecycle from collection to peer-reviewed publication, a gap the proposed framework explicitly targets.

III. Proposed System Design

In this section we present the developed platform [9]. Figure 1 illustrates our modular architecture comprising five layers:

1. Frontend Layer: React-based responsive UI with real-time WebSocket connections for collaborative editing.

2. API Layer: FastAPI asynchronous services handling user requests, paper management, and AI orchestration.

3. AI Integration Layer: Vendor-agnostic abstraction supporting four providers with unified interface.

4. NLP Pipeline: PDF parsing (PyPDF2), text extraction, and writing style feature computation.

[2]https://prowritingaid.com
[3]https://www.overleaf.com
[4]https://docs.google.com
[5]https://www.researchgate.net
[6]https://www.mendeley.com
[7]https://www.authorea.com
[8]https://www.openclinica.com
[9]https://github.com/ElisavetKanidou/Research-Chat-Platform

5. Data Layer: PostgreSQL database with JSON columns for flexible metadata storage (writing style features, personalization settings).

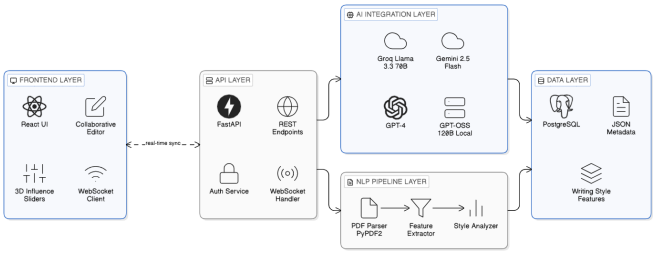

Fig. 1. System architecture of the proposed five-layer framework.

A. Three-Dimensional Personalization Model

1) Conceptual Framework: We model AI personalization as a three-dimensional influence space where users configure the relative importance of:

Lab Influence ($L \in [0, 10]$): Degree to which AI suggestions should reflect institutional research patterns. High lab influence ($L \geq 8$) prioritizes methodologies and terminology conventions from the user's research group publications.

Personal Influence ($P \in [0, 10]$): Degree to which AI should adapt to the individual author's writing style. High personal influence ($P \geq 8$) maintains consistency with the user's established vocabulary complexity, sentence structure, and citation patterns from prior publications.

Global Influence ($G \in [0, 10]$): Degree to which AI should incorporate field-wide best practices and contemporary terminology. High global influence ($G \geq 8$) aligns with recent literature trends and emerging methodologies in the medical specialty.

Unlike binary switches or single-dimensional sliders, our three-parameter model enables nuanced configurations: a junior researcher might set $(L = 9, P = 3, G = 7)$ to prioritize institutional conventions while learning field standards, whereas a senior investigator publishing in a new domain might choose $(L = 5, P = 8, G = 9)$ to maintain their established voice while incorporating cutting-edge practices.

2) Mathematical Formulation: The AI system prompt S is dynamically constructed as:

$$S = S_{\text{base}} + G_L(L) + G_P(P) + G_G(G) \tag{1}$$

where S_{base} is the foundational research assistant prompt, and G_L, G_P, G_G are guidance functions that append contextual instructions based on influence levels.

Lab Guidance Function $G_L(L)$:

$$G_L(L) = \begin{cases} \text{“Strongly prioritize lab patterns”} & L \geq 8 \\ \text{“Balance lab with alternatives”} & 5 \leq L < 8 \\ \text{“Focus on general best practices”} & L < 5 \end{cases} \tag{2}$$

Personal Guidance Function $G_P(P)$:

$$G_P(P) = \begin{cases} \text{“Mirror author's established style”} & P \geq 8 \\ \text{“Balance style with conventions”} & 5 \leq P < 8 \\ \text{“Prioritize field norms”} & P < 5 \end{cases} \tag{3}$$

Global Guidance Function $G_G(G)$:

$$G_G(G) = \begin{cases} \text{“Extensively use current trends”} & G \geq 8 \\ \text{“Balance trends with classics”} & 5 \leq G < 8 \\ \text{“Focus on proven methods”} & G < 5 \end{cases} \tag{4}$$

Additionally, the AI model temperature τ (controlling response creativity) adjusts based on global influence:

$$\tau = \tau_{\text{base}} + \frac{G}{20} \tag{5}$$

where $\tau_{\text{base}} = 0.6$. Higher global influence increases exploration of novel phrasing and contemporary methodologies ($\tau \in [0.6, 1.1]$).

3) Reference Context Integration: Our approach to context aggregation is conceptually related to Retrieval-Augmented Generation (RAG) frameworks [12], as it selectively integrates features from external reference papers into the generative prompt. When a user uploads a reference paper, our system:

1) **Categorizes** the paper as Lab, Personal, or Literature (Global) type
2) **Extracts** eight writing style features (Section III-C)
3) **Stores** features as JSON in database for rapid retrieval
4) **Aggregates** features during AI request construction based on influence levels

For example, with configuration $(L = 9, P = 4, G = 6)$ and 5 lab papers, 3 personal papers, 10 literature papers uploaded, the system constructs contexts:

$$C_L = \text{Aggregate features from 5 lab papers}$$
$$C_P = \text{Aggregate features from 3 personal papers}$$
$$C_G = \text{Aggregate features from 10 literature papers}$$

The final prompt includes C_L prominently (detailed statistics), C_P minimally (brief summary), and C_G moderately (key trends), proportional to the configured influence levels.

B. NLP Feature Extraction Pipeline

1) PDF Processing: Reference papers are uploaded as PDFs and processed via PyPDF2[10] to extract raw text while preserving section boundaries. Text undergoes cleaning (whitespace normalization, header/footer removal) before feature computation.

[10]https://pypdf2.readthedocs.io

2) Writing Style Features: We compute eight quantitative features characterizing writing patterns:

1. Average Sentence Length (μ_{sent}):

$$\mu_{\text{sent}} = \frac{1}{n}\sum_{i=1}^{n}|s_i| \tag{6}$$

where n is the number of sentences, $|s_i|$ is word count of sentence i. Medical papers typically range $\mu_{\text{sent}} \in [18, 28]$ words.

2. Vocabulary Complexity (Type-Token Ratio):

$$\text{TTR} = \frac{|V|}{N} \tag{7}$$

where V is the set of unique words, N is total word count. Higher TTR indicates diverse vocabulary.

3. Passive Voice Ratio (r_{passive}): Estimated by pattern matching for constructions: "(is—are—was—were—been) + past participle". Medical methods sections often exhibit $r_{\text{passive}} \in [0.3, 0.6]$.

4. Common Phrases: Top-10 bigrams by frequency, filtered to exclude stopword-only phrases. Examples: "clinical trial", "statistical significance", "patient outcomes".

5. Technical Terms: Extracted via two heuristics: (a) capitalized multi-word phrases (e.g., "Magnetic Resonance Imaging"), (b) complex words ≥ 10 characters with domain suffixes (-ology, -graphy, -tion). Ranked by frequency.

6. Citation Density (ρ_{cite}):

$$\rho_{\text{cite}} = \frac{\text{citation count}}{N} \times 1000 \tag{8}$$

Citations detected via patterns: (Author, Year) or [Number]. Reported as citations per 1000 words.

7. Section Structure: Ordered list of extracted section headings (Introduction, Methods, Results, Discussion, Conclusion) using regex patterns for numbered/capitalized headings.

8. Word Count Distribution: Mean and standard deviation of section lengths, indicating preferred granularity of content organization.

These features are stored per-paper in a JSON column, enabling rapid aggregation queries when constructing AI prompts.

C. Multi-Provider AI Integration

The platform uses an abstraction layer that can communicate with four different AI assistant models:

1. Groq Service (Llama 3.3 70B): - Fast inference (87ms average latency) - Recommended for general research writing

2. Gemini Service (Gemini 2.5 Flash): - Strong performance in biomedical and scientific question answering tasks

3. OpenAI Service (GPT-4): - Highest quality medical reasoning - Recommended for critical manuscript sections

4. GPT-OSS 120B: - On-premises deployment (with Ollama) - Zero external data transmission

Having in mind the sensitive research work which uses patient data and unpublished clinical research, we provide a local GPT-OSS model. Using this model the platform operates entirely within the institutional network without making any external API requests.

D. Real-Time Collaboration Infrastructure

1) WebSocket-Based Updates: The collaboration mechanism uses WebSocket connections which allow section-level change notifications to operate with 87ms average latency while showing collaborator online/offline status and supporting live comment threads and using last-write-wins timestamp-based automatic conflict resolution.

2) Role-Based Access Control: The system implements five hierarchical roles to manage access to shared papers. The **Owner** has full control and can delete the paper. The **Co-author** can edit content, invite collaborators, and manage settings. The **Editor** can edit content and add comments. The **Commenter** can add inline comments with view-only access to the content. Finally, the **Viewer** has read-only access. Invitations are sent via email, with instant access for existing platform users (2.3s average acceptance time) and a signup flow for new users.

IV. Evaluation

In this section, we present the evaluation of the proposed framework. The evaluation was designed and reported based on the guidelines of Runeson et al. for case studies [13]. The evaluation was conducted in two steps. The first step focuses on the usability and perceived usefulness of the system through a System Usability Scale (SUS) assessment. The second step is a domain-specific medical writing study, where we used the RAISE platform[11] to collect and manage medical research data and subsequently used the proposed framework to assist in drafting a research paper based on these data.

A. Step 1: System Usability Evaluation

Regarding the user study, we employed the System Usability Scale (SUS) [14], a 10-item questionnaire yielding a score from 0 to 100. Additionally, 30-minute semi-structured qualitative interviews were conducted with all participants.

1) Participants and Procedure: The usability evaluation was conducted with around 20 researchers recruited from different domains and experience levels. Each participant uploaded a number of reference papers (lab papers, personal publications, and key literature), configured personalization settings via the interactive sliders, completed writing tasks using the AI assistant (abstract drafting, literature review, methods description), filled out the SUS questionnaire, and participated in a 30-minute qualitative interview.

2) SUS Results: The SUS evaluation yielded a mean score of 80.78, which places the system at the boundary between "Good" and "Excellent" on the adjective rating scale, corresponding to a Grade A and "Acceptable" usability classification according to Bangor et al. [15]. This score exceeds the 68.0 industry average by approximately 12.8 points. At the item level, responses were predominantly positive across most dimensions. Items reflecting desire for frequent use, system consistency, ease of use, and confidence received ratings of 4 or 5 out of 5 from all or nearly all participants, with no

[11] https://raise-science.eu/

negative ratings recorded for these items. Integration of system functions was similarly well received. Greater variability was observed for items related to unnecessary complexity, learnability, and the need for technical support: three participants rated the system unnecessarily complex, two found it cumbersome, and one indicated they would need the support of a technical person to use it. These results suggest that while the system is perceived as highly usable and consistent overall, initial onboarding and configuration complexity remain the primary areas for improvement.

3) Qualitative Findings: The qualitative interviews corroborated the quantitative findings. Participants found the three-slider interface intuitive, with one noting that they could adjust lab influence for internal reports and increase global influence for journal submissions. The ability of the system to adapt to previously published vocabulary was highlighted as particularly impressive by participants with established publication records. Participants who handled patient data considered the local deployment option essential because it allowed AI support to function while keeping confidential data secure from external API transmission. The real-time collaboration feature received positive feedback because it allowed smooth co-authoring between different institutions.

The participants also identified some limitations. Consistent with the item-level variability observed for complexity-related statements, three participants found certain workflows unnecessarily complex and indicated they would prefer access to technical support during initial setup and configuration. Three additional participants found the initial reference paper upload process tedious (average 15 minutes for 5 papers). One participant experienced confusion when switching between AI providers due to differences in response quality.

B. Step 2: Medical Writing Study with RAISE Data

To evaluate the proposed framework in a realistic medical research scenario, we leveraged data collected through the RAISE platform (https://raise-science.eu/), a European funded research infrastructure. The RAISE platform supports the full lifecycle of research data, from collection, to utilize, and dissemination, making it an ideal testbed for a framework that aims to bridge the gap between data collection and scientific publication. Using the RAISE platform, we gathered clinical and research data from an ongoing study. The used dataset is tracking older adults with multiple chronic conditions after hospital discharge, focusing on frailty, cognition, quality of life, and the usability of digital health technologies during the transition to home.

In this study, 3 researchers from the initial participant pool collaborated to draft a paper using the proposed system. Prior to writing, the AI chat component was used to assess the quality and completeness of the RAISE dataset: researchers queried the system to identify missing variables, flag inconsistencies in data entry. This data quality verification step proves essential for the data management.

The researchers then uploaded relevant reference papers from their lab, personal publications, and current literature to configure the personalization settings. They used the AI assistant to draft specific sections of the paper (abstract, introduction, methods, and discussion), iterating on the output by adjusting the three-dimensional sliders to achieve the desired balance between institutional conventions, personal voice, and field-wide standards.

The findings of this study present three primary results. First, the three-dimensional personalization system demonstrated clear value across different manuscript sections: the methods section required high lab influence ($L = 9, P = 4, G = 5$) for consistency with established methodological standards, while the discussion section required higher global influence ($L = 5, P = 5, G = 9$) to connect the results with the broader academic literature. Second, the research team made extensive use of multi-author collaboration features, the role-based access control system established clear responsibility boundaries while permitting the lead author to maintain editing authority. Third, the local deployment option (GPT-OSS) was used exclusively for sections referencing patient-level data, in line with GDPR requirements, while all other writing activities were conducted through cloud providers. Taken together, these results demonstrate that the proposed framework supports not only the writing process but also the responsible secondary use of clinical research data in compliance with European data governance standards.

V. Discussion

A. Implications for Medical Research

Our three-dimensional model addresses a fundamental challenge in medical research: integrating institutional expertise, individual experience, and field evolution. Unlike end-to-end learned personalization that operates as a black box, our slider-based design maintains user agency, with all participants indicating they would like to use the system frequently in their research workflow.

Beyond writing quality, the framework contributes to a broader data lifecycle challenge. Clinical research datasets collected through platforms such as RAISE often remain underutilized after the primary study concludes. By lowering the barrier to manuscript preparation, the proposed system supports secondary data use and increases the probability that collected data will be transformed into published, FAIR-compliant scientific outputs. This aligns with European Health Data Space objectives and open science mandates increasingly required by funding agencies.

From an architectural perspective, institutional diversity in data governance policies necessitates flexible deployment models. Our abstraction layer future-proofs the platform as new LLMs emerge, since adding a new provider only requires implementing the unified interface.

B. Comparison with Existing Platforms

To position our framework against existing tools, Table II summarizes the comparison across key features. As it can

be observed, none of the existing platforms combines three-dimensional personalization, NLP-based style analysis, multi-provider AI support, and local deployment in a single open-source solution.

TABLE I
COMPARISON WITH EXISTING PLATFORMS

Feature	Research Platform	Grammarly	Notion AI	Overleaf
3D Personalization	✓	×	×	×
NLP Style Analysis	✓	×	×	×
Multi-Provider AI	✓	×	×	×
Local Deployment	✓	×	×	×
Real-time Collab	✓	×	✓	✓
Medical Focus	✓	×	×	×
Secondary Data Use	✓	×	×	×
Open Source	✓	×	×	×

C. Ethical Considerations

Our design emphasizes human oversight, requiring explicit user approval before incorporating any AI output. We recommend that users fact-check all AI-generated clinical claims against primary literature and disclose AI usage per emerging journal policies [16]. We also acknowledge that training data biases in foundation models may propagate to suggestions; our reference paper approach partially mitigates this by grounding recommendations in user-curated literature.

D. Limitations and Future Work

This study is subject to several limitations. First, the eight NLP features capture surface-level writing patterns but miss deeper rhetorical structures such as argumentative flow and hedging language. Future work will incorporate medical concept extraction using UMLS[12]. Second, the evaluation at a single institution limits generalizability, a broader multi-domain study is planned to address this constraint. Third, new users without publications cannot leverage personal influence, constituting a cold start problem that could be addressed through interactive style questionnaires or transfer learning from similar users. Fourth, feature aggregation consumes 800–1200 tokens per dimension, approaching context limits for extensive reference libraries, future work will explore dense vector representations for more efficient encoding.

VI. CONCLUSION

In this paper, we presented a three-dimensional personalization framework for AI-assisted medical research writing that enables dynamic control over Lab, Personal, and Global knowledge influences through interpretable sliders. The framework addresses a critical bottleneck in the health data lifecycle: the gap between clinical data collection and scientific dissemination. The multi-provider architecture supports four AI vendors, enabling institutions to select models based on data sensitivity requirements and GDPR compliance needs. The two-step evaluation demonstrated strong usability (SUS 80.78), and validated the framework in a realistic secondary data use scenario with data from the RAISE platform. The framework is released as open source to serve the research community.

[12]https://www.nlm.nih.gov/research/umls

ACKNOWLEDGMENT

This paper is a result of research conducted within the "MSc in Artificial Intelligence and Data Analytics" of the Department of Applied Informatics, University of Macedonia. The presentation of the paper is funded by the University of Macedonia Research Committee. This work was also conducted in the context of the RAISE Suite project, which has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101188337.

REFERENCES

[1] K. Dwan, C. Gamble, P. R. Williamson, J. J. Kirkham, and R. B. Group, "Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review," *PloS one*, vol. 8, no. 7, p. e66844, 2013.

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[3] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. D. S. Santos, P. E. Bourne *et al.*, "The fair guiding principles for scientific data management and stewardship: Comment," *Scientific data*, vol. 3, p. 160018, 2016.

[4] RAISE Consortium, "Raise platform: Research infrastructure for health innovation," https://raise-science.eu, 2024.

[5] Anthropic, "Claude 2: Improvements in thoughtfulness," https://www.anthropic.com, 2023, technical Report.

[6] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," vol. 35, 2022, pp. 27 730–27 744.

[7] R. Mir, B. Felbo, N. Obradovich, and I. Rahwan, "Evaluating style transfer for text," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 495–504.

[8] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?" in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2204–2213.

[9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," vol. 33, 2020, pp. 1877–1901.

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *Iclr*, vol. 1, no. 2, p. 3, 2022.

[11] P. A. Harris, R. Taylor, B. L. Minor, V. Elliott, M. Fernandez, L. O'Neal, L. McLeod, G. Delacqua, F. Delacqua, J. Kirby *et al.*, "The redcap consortium: building an international community of software platform partners," *Journal of biomedical informatics*, vol. 95, p. 103208, 2019.

[12] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," vol. 33, 2020, pp. 9459–9474.

[13] P. Runeson, M. Host, A. Rainer, and B. Regnell, *Case study research in software engineering: Guidelines and examples*. John Wiley & Sons, 2012.

[14] J. Brooke *et al.*, *SUS-A quick and dirty usability scale*. London, England, 1996, vol. 189, no. 194.

[15] A. Bangor, P. Kortum, and J. Miller, "Determining what individual sus scores mean: Adding an adjective rating scale," *Journal of usability studies*, vol. 4, no. 3, pp. 114–123, 2009.

[16] Nature Editorial, "Tools such as ChatGPT threaten transparent science; here are some principles for their use," *Nature*, vol. 613, p. 612, 2023.