TrueNorth: PERMA+4 and Conversational Agentic RAG to Optimize Long-Term STEM Engagement

Katja Crusius
Center for Information Systems and Technology
Claremont Graduate University
Claremont, CA, USA
0009-0002-1647-4860

Joan Puteri Zheng
Center for Information Systems and Technology
Claremont Graduate University
Claremont, CA, USA
0000-0001-5560-7975

Varsha Shetty
Center for Information Systems and Technology
Claremont Graduate University
Claremont, CA, USA
varsha-ravindra.shetty@cgu.edu

Sonia Baron
Division of Behavioral and Organizational Sciences
Claremont Graduate University
Claremont, CA, USA
0009-0003-7919-1013

Tracy Gaolese
Center for Information Systems and Technology
Claremont Graduate University
Claremont, CA, USA
tracy.gaolese@cgu.edu

William Wyatt
Division of Politics & Economics
Claremont Graduate University
Claremont, CA, USA
0009-0001-0710-1918

Anthony Dong
Department of Computer Science
San Francisco State University
San Francisco, CA, USA
adong@sfsu.edu

Windtell Souphavong
Department of Computer Science
San Francisco State University
San Francisco, CA, USA
wsouphavong@sfsu.edu

Yan Li
Center for Information Systems and Technology
Claremont Graduate University
Claremont, CA, USA
0000-0002-0415-0140

***Abstract*— The persistent "leaky pipeline" in STEM fields threatens workforce retention and technological progress. This paper applies Retrieval-Augmented Generation (RAG) to overcome limitations of large language models in delivering contextually accurate, evidence-based mentoring for STEM professionals. We present *TrueNorth*, a dialogue system that integrates knowledge retrieval with the PERMA+4 framework to enhance factual accuracy and mentorship effectiveness. TrueNorth employs an advanced agentic RAG architecture with nine autonomous modules, including optimized indexing, query refinement, semantic-aware chunking, and hallucination detection. The system draws on PERMA+4-aligned, peer-reviewed literature and best practices, leveraging Google Gemini for vector search and natural language generation within a PostgreSQL knowledge base. PERMA+4 extends well-being theory to organizational contexts by adding physical health, mindset, environment, and economic security to the traditional model. Evaluated on 116 test scenarios across 12 common STEM workplace challenges, TrueNorth delivered actionable, obtainable guidance, with 11 of 14 evaluation dimensions (79%) showing no significant category-dependent variation ($p > 0.05$), indicating robust and reliable performance across a variety of contexts. These results demonstrate TrueNorth's potential to support the multifaceted issue of early-career STEM development while upholding responsible AI principles.**

***Keywords*—conversational agents, PERMA+4 framework, retrieval-augmented generation, agentic AI, workforce retention**

I. Introduction

The United States faces growing demands for leadership in Science, Technology, Engineering, and Mathematics (STEM), particularly as global competition accelerates and domestic workforce demands intensify. While degree attainment across demographic groups has risen substantially in recent decades, this educational expansion has yet to translate into proportional participation or retention in STEM careers [1]. As documented in recent studies, the resulting gap not only undermines national competitiveness but also leaves critical high-skill positions unfilled, despite the presence of an educated talent pool [2].

Traditional mentoring programs have played an important role in addressing this challenge, with evidence highlighting their value [3]. Traditional mentoring programs, while valuable, often focus on large-scale institutional programs and mentorship initiatives, which, though well-intentioned, are frequently hampered by inefficiency, resource constraints, and overly complex organizational dynamics. There is a clear need for more accessible, low-cost solutions that empower individuals to take ownership of their career trajectories, reward merit, and deliver actionable, evidence-based guidance without the administrative burdens that can accompany traditional interventions [3].

Recent advances in large language models (LLMs) present promising opportunities to support existing mentorship structures through scalable, personalized guidance [4]. However, current solutions are rarely specific to the required domain expertise, empirical rigor, and appropriate context of individuals pioneering career trajectories in a professional and competitive STEM environment [5].

To address this gap, we present TrueNorth, a dialogue system designed to support early-career STEM professionals with context-specific, peer-reviewed guidance grounded in proven principles in organizational behavior. Built on a Retrieval-Augmented Generation (RAG) architecture, TrueNorth integrates targeted knowledge retrieval with natural language generation, ensuring both relevance and factual integrity. Its knowledge base is structured around the PERMA+4 framework [6][7], an evidence-based model of workplace well-being and performance, ensuring that guidance promotes both individual development and organizational excellence. By embedding PERMA+4 principles into the knowledge base and retrieval strategy, our approach ensures that the generated content addresses both professional development and workplace well-being. We further introduce an agentic RAG architecture composed of nine specialized AI modules, including semantic-aware chunking, hallucination detection, and optimized vector indexing, that is built to improve the factual accuracy and relevance of system outputs. By evaluating on 116 test scenarios reflecting common workplace challenges occurring in STEM fields, we demonstrate that our system delivers actionable and psychologically grounded guidance.

Implementation details and reproducible code can be found at: https://github.com/jovalie/TrueNorth-AIxHEART-2025.

II. BACKGROUND

The design of TrueNorth is informed by two areas in the existing literature. First, we review research on the leaky pipeline in STEM, focusing on the most common structural and psychological barriers experienced by early-career professionals in STEM. Second, we look at the PERMA+4 framework, which builds on the established PERMA framework from psychology for promoting well-being and long-term engagement in organizational settings.

A. The Leaky Pipeline Phenomenon

The term "leaky pipeline" refers to the progressive attrition of individuals from STEM fields at key transition points, including both postsecondary education and the early stages of professional careers, resulting in the attrition of our most educated workers in a high-need sector [8]. National workforce data continues to show that retention in STEM professions varies significantly across American populations, despite high performance in related university coursework [2][9].

Longitudinal studies reveal that preventable factors like lack of mentorship, high sense of alienation, and a diminished scientific identity heavily contribute to the attrition of early-career STEM professionals. For instance, [2] demonstrates that students who experiences quality mentorship and research engagement showed significantly higher persistence in STEM up to four years post-graduation [9]. These findings emphasize that mentorship is essential across STEM fields, especially for professionals that are pioneering transformative technologies.

B. The PERMA+4 Framework for Workplace Flourishing

To support evidence-based mentoring in STEM, our system is grounded in the PERMA+4 framework which is an extension of Seligman's original PERMA model of well-being [6]. PERMA+4 captures give core dimensions of human flourishing, Positive Emotion, Engagement, Relationships, Meaning, and Accomplishment. This is expanded by PERMA+4, adding Physical Health, Mindset, Environment, and Economic Security which makes it highly applicable to organizational and professional contexts [7].

This expanded framework addresses the complex, interdependent factors that influence young professionals' well-being and career satisfaction in STEM fields. By integrating PERMA+4, our dialogue system can guide users not only on professional achievement but also help with creating sustainable and fulfilling work conditions. PERMA+4 case studies of impactful organizational behavior in a professional and competitive STEM context can be retrieved with our knowledge base, allowing that retrieved and generated content aligns with empirically supported workplace psychology insights. We curated PERMA+4 aligned literature from peer-reviewed sources to serve as the knowledge base, allowing our system to deliver nuanced, targeted guidance in real time to support young professionals.

C. Design Principles

To ensure that TrueNorth effectively supports early-career STEM professionals, we grounded its interaction design in the meta-requirements and design principles proposed by [10]. These principles offer a guideline for creating stereo-type neutral, intelligent interfaces that foster trust, inclusion and empowerment – key values that align with the PERMA+4 framework integrated into our system. These meta-requirements consolidate into three design principles optimized support pioneering STEM professionals.

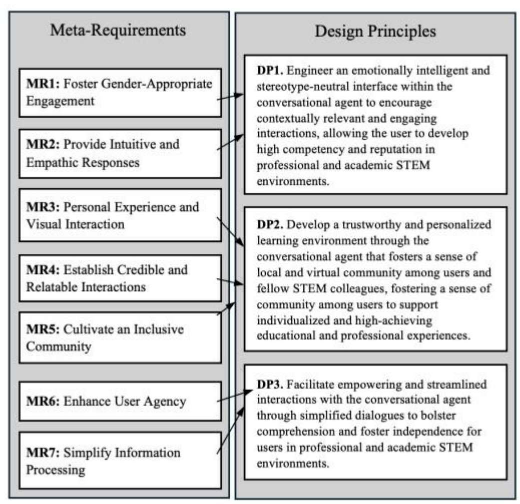

Figure 1: Meta-requirements and design principles for promoting retention in STEM as inspired by [10].

III. TRUENORTH SYSTEM OVERVIEW

Retrieval-Augmented Generation combines the generative capabilities of pre-trained language models with dynamic information retrieval from external knowledge bases [11]. Unlike traditional LLMs that rely solely on parametric knowledge learned during training, RAG systems augment the generation process by retrieving relevant documents or passages from curated databases at inference time. This approach is particularly valuable for applications requiring up-to-date information, domain-specific expertise, or factual grounding that may not be present in the model's training data.

Google Gemini 2.0 Flash was used for language generation using parameters `{temperature = 1.0, top-p = 0.95, top-k = 64, and candidate-count=1}`. The RAG pipeline consists of three primary components: (1) a retrieval system that identifies relevant documents based on semantic similarity to the input query, (2) a knowledge base containing domain-specific information, and (3) a generative model that synthesizes responses using both the original query and retrieved context. This architecture enables the system to provide more accurate, contextually relevant responses while maintaining the flexibility and conversational capabilities of modern LLMs.

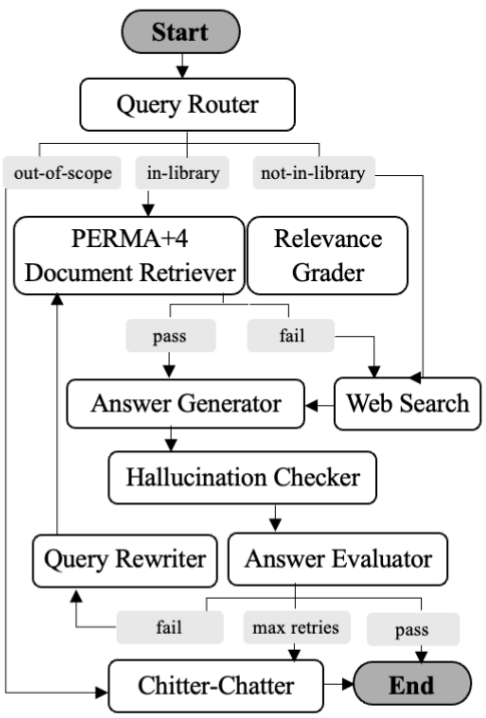

Figure 2: Multi-agent RAG Dialogue System with PERMA+4 Knowledge Base. Conversation is informed by from priority documents stored as vectorized embeddings and Tavily Web Search API.

A. Knowledge Base and Vectorization

The system's knowledge foundation consists of 32 carefully selected citations spanning documents spanning positive psychology, emotion regulation, conflict resolution, leadership theory, and empirical research on PERMA+4 centered evaluations of STEM early career professionals. Each document underwent systematic preprocessing, including header-based chunking to preserve contextual integrity while optimizing for retrieval accuracy.

Document vectorization employs Google Gemini's `text-embedding-004` model, chosen for its current competitive and cost-effective performance in capturing semantic relationships within academic and professional contexts. The resulting embeddings are stored in PostgreSQL with `pgvector`, enabling efficient similarity searches while maintaining data security through internally restricted VPN access requirements.

B. Multi-Agent Architecture

Our conversational agent employs a sophisticated nine-agent architecture built using LangChain for orchestration and LangGraph for managing conversational flows. Each agent operates with specialized prompts, distinct objectives, and targeted API access to ensure optimal performance across different query types and user needs. Full agent prompts including response formatting details can be found in GitHub repository.

- **Query Router:** Serves as the system's entry point, prompted to analyze incoming user queries to determine the appropriate processing pathway. This component evaluates query intent and context to route requests to either internal knowledge retrieval, external web search via the Tavily API, or general conversational handling through the Chitter-Chatter agent. The router's decision-making process is critical for system efficiency, as it prevents unnecessary computational overhead while ensuring users receive the most relevant information sources.
- **Document Retriever**: Performs semantic similarity searches on the vectorized knowledge database stored in PostgreSQL with `pgvector` extensions. Using Gemini embeddings, this component identifies the most contextually relevant document chunks from the curated knowledge base of 32 peer-reviewed sources.
- **Answer Generator:** Synthesizes final responses using Gemini 2.0 Flash, prompted to integrate retrieved knowledge, conversational history, and carefully crafted system prompts. The generation process incorporates conversation memory to maintain continuity across multi-turn dialogues, essential for building trust in mentoring relationships.
- **Relevance Grader**: Evaluates content obtained by Document Retriever using maximal marginal relevance (MMR) (Eq. 1) and reciprocal rank fusion (RRF) (Eq. 2), assigning confidence scores to ensure only the most pertinent information informs responses.

$$MMR = arg \max_{d_i \in D \setminus R} [\lambda \cdot Sim_1(d_i, q) - (1-\lambda) \cdot \max_{d_j \in R} Sim_2(d_i, d_j)] \quad (1)$$

D is the set of all candidate documents, R is the set of already selected documents, q is the query, Sim_1 is the similarity function between a document and a query, Sim_2 is the function between two documents, and d_i and d_j are documents in D and R respectively.

$$RRF(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)} \quad (2)$$

D is the set of all documents, R is the set of rankings, k is a constant set to 60.

- **Hallucination Checker:** Cross-references generated outputs against source documents and validates all included URLs through HTTP HEAD requests, preventing the propagation of inaccurate or unobtainable information.
- **Answer Evaluator:** Assesses response quality for conversational agent effectiveness. This agent evaluates generated responses across our three design principles as informed by [10]. This component is essential for maintaining consistent service quality and identifying potential bias or effectiveness issues across different user demographics and query types.
- **Query Rewriter**: When initial system processing fails due to query ambiguity, this agent provides a second opportunity for successful interaction by presenting the reformulated query to the retrieval pipeline.
- **Chitter-Chatter Agent**: Specializes in maintaining conversational flow while gently guiding users toward more substantive mentoring topics when appropriate.

The nine agents operate within an orchestrated workflow managed by LangGraph, enabling stateful conversation handling and dynamic routing based on query characteristics and system state. This coordination ensures that each agent's specialized capabilities are leveraged optimally while maintaining conversational coherence and minimizing response latency. The modular architecture allows for independent optimization of individual agents while preserving the system's overall functionality and user experience quality.

C. User Interface

The system interface is implemented using Streamlit, chosen for its rapid deployment capabilities and native support for both web and mobile platforms. While Streamlit provided efficient development iteration, the team identified limitations in creating polished, intuitive user experiences, prompting consideration of alternative technologies for future implementations focused on enhanced accessibility and user experience design.

The interface design prioritizes accessibility and emotional safety, incorporating clear disclaimers about the system's role as a supplement to, rather than replacement for, professional counseling or legal advice. Response presentation emphasizes readability and emotional tone appropriate for high stakes mentoring contexts.

IV. EVALUATION

This section outlines the initial evaluation framework employed to assess effectiveness of a conversational mentoring agent for early career STEM professionals.

A. Evaluation Framework Design

The evaluation methodology was informed by design principles for conversational agents serving early-career populations in technology education [10]. We adapted their seven-dimensional framework to address the unique requirements of professional and academic mentoring contexts, ensuring that our assessment captured both technical performance and user-centered effectiveness.

B. Test Dataset

We developed a evaluation dataset comprising 116 test questions spanning 14 themes identified by literature review performed in [12] as frequently common challenges experienced by early-career professionals in STEM. As shown in Table 1, each test question included a realistic user prompt, an optimal response benchmark, and categorical classification for consistent evaluation of performance across common themes experienced in STEM professional environments. Prompting system for synthetic data can be found in the GitHub repository referenced in the introduction. An IRB review to have TrueNorth tested by actual users is currently pending.

Category	Description
Professional Development & Career Advancement	Opportunities and challenges in skill growth, promotions, and long-term planning
Workplace Dynamics & Communication	Scenarios involving interruption patterns, credit attribution, and assertiveness challenges
Work-Life Balance & Family Expectations	Queries related to caregiving responsibilities, parental leave transitions, and family pressure
Mentorship & Sponsorship	Questions about accessing guidance, building professional networks, and finding mentors
Verbal Aggression	Scenarios involving direct/indirect discrimination, stereotype threat, and environmental hostility
Leadership & Authority	Challenges around establishing credibility, managing teams, and executive presence
Technical Competence	Situations where expertise is undermined or questioned
Isolation & Alienation	Issues of social integration, imposter syndrome, and community building
Compensation & Recognition	Questions about salary negotiation, performance evaluation, and visibility
HR & Reporting Challenges	Scenarios involving formal complaint processes and institutional support
Motivation & Burnout	Queries addressing emotional resilience, goal setting, and sustainability
Gendered Culture	Gender-centric themes affecting participation in STEM environments

Table 1: Common challenges experienced by early career STEM professionals identified in literature review[12]

C. Multi-Dimentional Assessment Criteria

Our evaluation integrated seven core conversational agent metrics from [10], enabling systematic evaluation of both fundamental conversational quality and domain-specific effectiveness for supporting early career professionals navigating professional and academic STEM environments.

Metric	Description
Usefulness	Measured the extent to which responses provided actionable, relevant guidance that could meaningfully address the user's stated concern.
Goal Facilitation	Assessed how effectively responses supported users in achieving their underlying objectives, such as career advancement, conflict resolution, or development.
Relatability	Evaluated whether responses demonstrated understanding of challenges faced by actual professionals in STEM, with cultural context and professional dynamics acknowledged.
Trustworthiness	Measured user confidence in the accuracy, credibility, and reliability of information, with appropriate citations and acknowledgment of limitations.
Accessibility	Assessed the clarity, comprehensibility, and inclusivity of language, ensuring suitability for varying educational and professional backgrounds.
Human-likeness	Measured the naturalness and conversational quality of responses, including empathy, tone, and emotional intelligence.
Attractivity	Evaluated the overall appeal and engagement of responses, including visual presentation and likelihood of continued user interaction.

Table 2: Evaluation metrics for our dialogue system, defined in [10].

D. Evaluation Procedure

We employed Anthropic's `Claude-3.5-haiku` model as an automated judge to evaluate the conversational agent's responses across all seven dimensions using a 5-point Likert scale (1 = highly disagree, 5 = highly agree). The AI judge was provided with detailed rubrics for each evaluation criterion, example responses at different quality levels, and specific instructions to maintain consistency across assessments.

The AI judge was provided with detailed rubrics for each evaluation criterion, example responses at different quality levels, and specific instructions to maintain consistency across assessments. Each of the 116 generated responses was evaluated independently across all fourteen dimensions, enabling systematic, scalable evaluation of 1,568 individual assessments while maintaining consistency in assessment criteria application across the expanded test set.

V. RESULTS

This section presents the evaluation results of our RAG conversational agent across fourteen key dimensions using AI-as-a-judge methodology applied to 116 test cases covering twelve distinct challenge categories commonly experienced by early-career STEM professionals.

A. Performance Analysis

Our conversational agent demonstrated strong overall performance across all evaluation dimensions, with mean scores consistently exceeding 4.0 on the 5-point Likert scale. The system achieved its highest performance in Accessibility (mean = 4.7, SD = 0.3) and Goal Facilitation (mean = 4.5, SD = 0.4), indicating exceptional ability to communicate in clear, comprehensible language and understand user objectives across varying challenge scenarios.

Trustworthiness scored highly (mean = 4.4, SD = 0.5), reflecting the system's success in grounding responses in authoritative sources from psychology, leadership, and gender studies research while maintaining credible, evidence-based guidance. The Usefulness dimension achieved a mean score of 4.3 (SD = 0.6), demonstrating that users consistently received actionable, relevant guidance tailored to their specific concerns.

Metric	Avg	Std	χ^2	df	p
Usefulness	4.3	0.6	21.95	26	0.692
Goal Facilitation	4.5	0.4	36.21	29	0.598
Relatability	4.0	0.8	54.65	39	0.049*
Trustworthiness	4.4	0.5	49.31	39	0.125
Accessibility	4.7	0.3	16.34	13	0.231
Human-likeness	4.1	0.7	31.50	26	0.210
Attractivity	3.9	0.8	26.49	26	0.437
Gender-Consciousness (MR1)	4.0	0.7	46.32	39	0.196
Empathic Intuition (MR2)	4.4	0.5	36.76	39	0.572
Personal Visual Engagement (MR3)	3.8	0.9	69.44	39	0.002**
Credibility & Relatability (MR4)	4.2	0.6	35.68	39	0.622
Inclusive Community (MR5)	3.9	0.8	59.98	39	0.021*
User Agency (MR6)	4.3	0.6	42.59	39	0.319
Cognitive Simplicity (MR7)	4.1	0.7	34.79	39	0.662

*Table 3: Performance metrics across 14 evaluation dimensions for the TrueNorth conversational agent, assessed using AI-as-a-judge methodology on 116 test cases covering 12 challenge categories commonly experienced by early-career STEM professionals. All metrics evaluated on 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). Chi-squared tests examine category-dependent performance variation across challenge types. Evaluation conducted using Claude-3.5-Haiku as AI judge with structured rubrics for each dimension. *p < 0.05, **p < 0.01*

B. Chi-Squared Analysis

The chi-squared analysis across 14 evaluation dimensions revealed that TrueNorth demonstrates remarkably consistent performance across most challenge categories, with 11 of 14 metrics (79%) showing no significant category-dependent variation ($p > 0.05$), indicating robust, reliable responses regardless of the specific type of early-pioneering STEM challenge presented.

C. Category-Specific Performance

Examination of mean scores across challenge categories revealed meaningful performance patterns that align with the system's knowledge base strengths and limitations. Of the 14 evaluation metrics analyzed, three showed statistically significant associations with challenge categories.

1. *Highest Performance Categories*

- **Accessibility** (M = 4.7, SD = 0.3, $\chi^2 = 16.34$, p = 0.231): Demonstrated the strongest and most consistent performance with no significant category variation, indicating reliable ease of interaction across all challenges.
- **Goal Facilitation** (M = 4.5, SD = 0.4, $\chi^2 = 36.21$, p = 0.598): Consistently strong support for promoting user objectives without significant category-dependent variation.
- **Empathic Intuition (MR2)** (M = 4.4, SD = 0.5, $\chi^2 = 36.76$, p = 0.572): Maintained high empathetic responsiveness across categories.

2. *Features for Further Optimiaztion*

- **Personal Visual Engagement (MR3)** (M = 3.8, SD = 0.9, $\chi^2 = 69.44$, p = 0.002*): While showing moderate performance, this metric shows both the lowest mean score among meta-requirements and the highest performance variation across challenge categories. The significant chi-squared result indicates that TrueNorth's ability to create personalized, contextually relevant interactions is inconsistent and would benefit from further adaptation to broader user contexts.
- **Inclusive Community (MR5)** (M = 3.9, SD = 0.8, $\chi^2 = 59.98$, p = 0.021**): Despite achieving a moderate mean score, the significant category variation indicates that TrueNorth's inclusive community-building capabilities vary across different categories of STEM challenges.
- **Identification/Relatability** (M = 4.0, SD = 0.8, $\chi^2 = 54.65$, p = 0.049*): While achieving an acceptable mean score, the significant category-dependent variation suggests that TrueNorth struggles to maintain consistent relatability per challenge theme, particularly for culturally sensitive or domain-specific concerns where personalized understanding is crucial, showing that further adaptation to broader user contexts is necessary.

VI. Discussion and Conclusion

The mixed statistical significance in category-dependent performance variation (with only 3 of 14 metrics showing significant effects) underscores the critical need for human validation in conversational AI evaluation. Future work should prioritize human-in-the-loop approaches to ensure that automated assessments accurately reflect genuine user experience and system effectiveness.

The prototype demonstrates how retrieval-augmented dialogue systems can integrate PERMA+4 factors as concurrent considerations to act as an interactive interface for positive STEM trajectory. This dialogue system provides reliable guidance while maintaining the empathetic, conversational qualities essential for effective mentoring relationships, where accuracy and individual context are equally critical.

TrueNorth makes a knowledge contribution by operationalizing the PERMA+4 framework with a RAG-based architecture and offering a replicable model for integrating psychological well-being principles into AI-driven career support systems. From a practical perspective, TrueNorth delivers a mentoring system tailored to the technically nuanced challenges that early-career STEM professionals face, closing existing gaps and promoting sustainable workforce retention.

Acknowledgments

This material is based upon work supported by the U.S. National Science Foundation under Grant No. 2313998, the Claremont Graduate University Travel & Material Award, the Computing Research Association Undergraduate Research to PhD Program (UR2PhD), and the Grace Berry Award for Graduate Women from Claremont Colleges Intercollegiate Feminist Center for Teaching, Research and Engagement. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. National Science Foundation.

References

[1] L. Malcom and S. Malcom, "The Double Bind: The Next Generation," *Harvard Educational Review*, vol. 81, no. 2, pp. 162–172, Jun. 2011, doi: 10.17763/haer.81.2.a84201x508406327.

[2] A. Almukhambetova, D. H. Torrano, and A. Nam, "Fixing the Leaky Pipeline for Talented Women in STEM," *Int J of Sci and Math Educ*, vol. 21, no. 1, pp. 305–324, Jan. 2023, doi: 10.1007/s10763-021-10239-1.

[3] A. Byars-Winston and M. L. Dahlberg, Eds., *The Science of Effective Mentorship in STEMM*. Washington, D.C.: National Academies Press, 2019. doi: 10.17226/25568.

[4] M. M. Rashid *et al.*, "Humanizing AI in Education: A Readability Comparison of LLM and Human-Created Educational Content," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 68, no. 1, pp. 596–603, Sep. 2024, doi: 10.1177/10711813241261689.

[5] F. Beroíza-Valenzuela and N. Salas-Guzmán, "Redefining academic trajectories: A comprehensive analysis of the factors and impacts of the gender gap in STEM higher education," *IJSHE*, vol. 26, no. 4, pp. 928–947, Apr. 2025, doi: 10.1108/ijshe-02-2024-0129.

[6] S. I. Donaldson, L. E. van Zyl, and S. I. Donaldson, "PERMA+4: A Framework for Work-Related Wellbeing, Performance and Positive Organizational Psychology 2.0," *Front. Psychol.*, vol. 12, Jan. 2022, doi: 10.3389/fpsyg.2021.817244.

[7] V. Cabrera and S. I. and Donaldson, "PERMA to PERMA+4 building blocks of well-being: A systematic review of the empirical literature," *The Journal of Positive Psychology*, vol. 19, no. 3, pp. 510–529, May 2024, doi: 10.1080/17439760.2023.2208099.

[8] D. A. Barr, M. E. Gonzalez, and S. F. Wanat, "The leaky pipeline: factors associated with early decline in interest in premedical studies among underrepresented minority undergraduate students," *Acad Med*, vol. 83, no. 5, pp. 503–511, May 2008, doi: 10.1097/ACM.0b013e31816bda16.

[9] M. Estrada, P. R. Hernandez, and P. W. Schultz, "A Longitudinal Study of How Quality Mentorship and Research Experience Integrate Underrepresented Minorities into STEM Careers," *LSE*, vol. 17, no. 1, p. ar9, Mar. 2018, doi: 10.1187/cbe.17-04-0066.

[10] D. Stattkus, L. Göritz, K.-M. Illgen, J. Beinke, and O. Thomas, *A Blueprint for Breaking Stereotypes: Design Principles for Conversational Agents to Empower Girls in Computer Science Education*. 2024.

[11] A. Singh, A. Ehtesham, S. Kumar, and T. T. Khoei, "Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG," Feb. 04, 2025, *arXiv*: arXiv:2501.09136. doi: 10.48550/arXiv.2501.09136.

[12] Y. Makarem and J. Wang, "Career experiences of women in science, technology, engineering, and mathematics fields: A systematic literature review," *Human Resource Development Quarterly*, vol. 31, no. 1, pp. 91–111, 2020, doi: 10.1002/hrdq.21380.