IFN712 Research Project Proposal Form

(Submitted to [y.feng@qut.edu.au](mailto:y.feng@qut.edu.au) by 30 June 2025)

|  |  |
| --- | --- |
| Project agency (school, industry) | School of Computer Science, School of Information Systems |
| Industry supervisor and contact emails | Lingfeng Ye, l3.ye@qut.edu.au |
| Academic Supervisor name(s) and contact emails | Yanming Feng [y.feng@qut.edu.au](mailto:y.feng@qut.edu.au);  Lingfeng Ye, l3.ye@qut.edu.au |
| Information Technology major(s) | Software Development, Computer Science and Data Science |
| Project title | AI-Driven Project-Student Matching under Data Scarcity and Privacy Constraints |
| Brief description of the research problem, aims, method and expected outputs (100~200 words) | **This project explores methodological foundations for matching research projects with suitable researchers without relying on publicly-available, fully-labelled datasets. The research addresses fundamental challenges in academic and industry environments where traditional matching approaches fail due to inherent constraints and uncertainties. The research environment is characterised by uncertain data volume, where the number and richness of project descriptions and researcher profiles fluctuate widely across organisations and time periods. Additionally, no reliable ground-truth labels exist, as there is no large corpus of "successful match vs. failed match" outcomes available for training or benchmarking. Strict privacy and security constraints further complicate the environment, as project briefs may contain proprietary ideas and researcher profiles may hold personal data that cannot leave institutional boundaries.**  **This investigation systematically examines four orthogonal design dimensions to develop robust methodological approaches for project-researcher matching under these constraints. The central challenge involves researching, prototyping and evaluating design choices across these dimensions when ground-truth is missing and usable data may be scarce or siloed.**  **Research Objectives:**  The investigation addresses four design dimensions through specific methodological developments:   * **Canonical Matching Process**: Determine optimal sequences and combinations of processing steps (preprocessing, candidate retrieval, re-ranking, rule-based filtering, explanation) through comparative analysis of pipeline architectures, ablation studies on component contributions, and development of quality metrics that operate without ground-truth labels. * **Adaptive Pipeline Design**: Examine the effects of varying run-time autonomy levels from fixed pipelines to self-adaptive agents through implementation across the autonomy spectrum, development of self-monitoring mechanisms, and analysis of consistency versus adaptability trade-offs. * **Multi-Tool Orchestration**: Establish criteria and frameworks for selecting and governing retrieval, ranking, and reasoning tools (vector search, BM25, cross-encoders, LLM scoring, knowledge graphs) through development of tool selection criteria, comparative evaluation across data regimes, and meta-learning approaches for tool orchestration. * **Privacy-Preserving Matching**: Integrate security techniques (differential privacy, federated computation, access-controlled knowledge stores) while maintaining performance through implementation of privacy-preserving techniques, development of privacy-utility trade-off metrics, and security analysis under adversarial conditions.   The central challenge involves researching, prototyping and evaluating design choices across these dimensions when ground-truth is missing and usable data may be scarce or siloed. |
| Key words | Retrieval-Augmented Generation (RAG), Agentic AI system, Data Privacy, Information Retrieval, Semantic Vector Search, Knowledge Graph, Model Context Protocol |
| Answerable research questions for 3-5 students (optional) | 1. In the absence of labelled outcomes, which sequence and combination of canonical steps—e.g., preprocessing, candidate retrieval, re-ranking, rule-based filtering, explanation—yields the best trade-off between matching quality, interpretability, and computational cost under variable data availability? 2. How does increasing the degree of run-time autonomy—from a fixed pipeline through a master-planned pipeline to highly self-adaptive agents—affect matching accuracy, output consistency, and failure recovery, and which constraint & monitoring mechanisms are required to keep an adaptive pipeline reliable when no ground-truth feedback is available? 3. What criteria and decision framework enable an agent to select and orchestrate retrieval, ranking, reasoning, and rule-based tools (vector search, BM25, cross-encoders, LLM scoring, knowledge graphs, etc.) so that precision and recall are maximised across different data-volume regimes while operating without supervised labels? 4. Which privacy-preserving and security techniques (e.g., differential privacy, federated or split computation, access-controlled knowledge stores) can be integrated into the matching pipeline with minimal degradation of performance, and how can their effectiveness and side-effects be evaluated when explicit ground-truth labels are unavailable? |
| 3-5 key references (optional) and environment scan sources | 1. Retrieval-Augmented Generation for Large Language Models: A Survey, <https://doi.org/10.48550/arXiv.2312.10997> 2. ReAct: Synergizing Reasoning and Acting in Language Models, <https://doi.org/10.48550/arXiv.2210.03629> 3. A Survey on Knowledge Graphs: Representation, Acquisition and Applications, <https://doi.org/10.48550/arXiv.2002.00388> 4. Differentially Private Synthetic Data via Foundation Model APIs 2: Text, <https://doi.org/10.48550/arXiv.2403.01749> |
| Required major of studies, skills, knowledge, and speciality | The project is suitable for students in software development and computer science major. Each student will focus the questions and aspects relevant to own major. Students may develop the skills in the following areas   * **Natural Language Processing (NLP)** * **Data Engineering & Management** * **AI Security & Privacy Engineering** * Accessing LLM API Services |
| **Industry-based project: Student IP Agreement.** This is the IP model agreed between the parties. Please note that it is QUT policy that where possible students should be allowed to keep their IP. If students are asked to assign their work then please **provide a brief rationale** as additional permissions are needed by QUT to approve. | Project IP vests in the student with a license back to Industry Partner **(licence)**  OR  Project IP vests in the Industry Partner with a licence back to the student **(assignment)**  OR  Academic project |
| Number of students | 4 |
| The message from supervisor(s) about the acceptance for this project |  |
| Student name(s)  (Print your name and submit this form by the end of Week 2) |  |
| Date |  |
| Remarks on conditions of offer | The supervising team will shortlist the candidates after their application. |