# Fall 2025
# Generative Information Retrieval

**Final Project**

Sep 16, 2025

## Final Project

### Introduction

In the final project, you will work in groups of **3 ~ 4 students** to apply methods learned in the class to a real-world problem that you are interested in. In this final project, the topic scope includes but is not limited to: (1) Personalization and Recommendations in Search, (2) Agentic IR, (3) Multimodal Retrieval, and (4) Domain-Specialized RAG. All of them have been popular and highly focused topics in recent years, especially when LLMs are introduced. The detailed descriptions of each topic are listed below ("Topics Introduction" section) for your reference and inspiration of your project.

### Grading

The final project score is composed of four parts (total 40%):

- **Initial Project Proposal Presentation (5%)**
- **Mid-term Project Results Presentation (10%)**
- **Final Project Report (15%)**
- **Final Project Presentation (10%)**

### Requirements

- Initial Project Proposal Presentation (**10/7 in class**)
    - You will present what you plan to do for your project in **3 minutes**. It's totally fine if your project eventually evolves into something else; the proposal is just to help you to clearly frame your project. However, your presentation should lay out a sensible initial plan. Below is what you need to cover:
    - **Introduction** - Brief overview of your problem. Why might this problem be important? What IR tasks will you address?
    - **Dataset** - Description of the dataset you plan to use – where you will collect the dataset from, the domain of the dataset, the size of the dataset, etc.
    - **Evaluation Methods** - Specify at least one well-defined, numerical, automatic evaluation metric you will use for quantitative evaluation. If you have any

particular ideas about the qualitative evaluation you will do, you can describe that too.

- Mid-term Project Results Presentation (**11/11 in class**)
    - You will present the related work you have surveyed, the methods you have implemented, and your preliminary results at the project in **5 minutes**. Below is what you need to cover:
    - **Introduction** - Brief overview of your problem. Why might this problem be important? What IR tasks will you address?
    - **Literature Review / Related work** - Description of other work/papers you've found that are related to your task. Just mentioning a paper is not sufficient; you should at least go into brief detail about what kind of approach they are using/how it relates to your work if it's not immediately clear. Please also mention why your work relates to or differs from these related works.
    - **Dataset** - Description of data you are using - the size of the dataset, distribution of classes, any preprocessing you needed to do
    - **Main Approach** - Propose a model, an algorithm, or a framework for tackling your task. You should describe the model and algorithm in detail and use a concrete example to demonstrate how it works. Don't describe methods in general; describe precisely how they apply to your problem (what are the inputs/outputs, variables, factors, states, etc.)?
    - **Evaluation Methods** - Specify at least one well-defined, numerical, automatic evaluation metric you will use for quantitative evaluation. If you have any particular ideas about the qualitative evaluation you will do, you can describe that too.
    - **Preliminary Results** - Describe the performance and your preliminary results analysis to see whether the results are expected. If the results do not meet your expectations, think about what you can do to enhance the performance and present your next step plan.
- Final Project Reports (**submission deadline: 12/15 23:59**)
    - You have to submit a **final project report** to showcase your project results. We note that **making a demo for your work is optional**. Below is a full description of what you should include in your final project report. We will grade your score based on these sections.
    - **Introduction** - Brief overview of your problem. Why might this problem be important? What IR tasks will you address?
    - **Literature Review / Related work** - Description of other work/papers you've found that are related to your task. Just mentioning a paper is not sufficient; you should at least go into brief detail about what kind of approach they are using/how it relates to your work if it's not immediately clear. Please also mention why your work relates to or differs from these related works.
    - **Dataset** - Description of data you are using - the size of the dataset, distribution of classes, any preprocessing you needed to do
    - **Baseline** - Description and implementation of your baseline.

    - **Main Approach** - Propose a model, an algorithm, or a framework for tackling your task. You should describe the model and algorithm in detail and use a concrete example to demonstrate how it works. Don't describe methods in general; describe precisely how they apply to your problem (what are the inputs/outputs, variables, factors, states, etc.)?
    - **Evaluation Metric** - Please include what metrics, both qualitative and quantitative, you are using to evaluate the success of your problem. If relevant please include equations to describe your metrics.
    - **Results & Analysis** - Please include the performance of your baseline as well as the performance of your main approach so far and any experiments that you have run. If your results are creative and can't find a proper baseline, then you can analyze how you get the results you want. To sum up, include an analysis of your results, and how this might inform your next steps in fine-tuning your main approach. The analysis is very important, and it requires you to think about what your results might mean.
    - **Error Analysis** - Describe a few experiments that you ran that show the properties (both pros and cons) of your system. Analyze the data and show either graphs or tables to illustrate your point. What's the take-away message? Were there any surprises? Use these experiments in the error analysis to describe potential errors in the method and why they may have occurred.
    - **Future Work** - This section can be short, but please include some ideas about how you could improve your model if you had more time. This can also include any challenges you're running into and how you might fix them.
    - **Code** - Please include a link to your Github/Bitbucket/etc. Your repo should include an overview of the task, prerequisite (your coding environment, package version (e.g., requirements.txt in Python)), usage, hyperparameters you set, experiment results, and so on.
    - **Contribution of each member** - Please include the contribution of each member with **proportions**. We understand the condition that some members may fail to contribute to this project; thus, we will adjust your score if the contributions are significantly unequal. Feel free to let us know if you have any concerns about this part.
    - **References** - Please include a reference section with properly formatted citations (any format of your choice).
- Final Project Presentation (**12/23 in class**)
    - You will present what you've done for your project in **5 minutes**. Note that you don't have much time to describe all the details in the presentation, and TA will conduct a strict time control. Below is what you need to cover:
    - **Introduction** - Brief overview of your problem. Why might this problem be important? What IR tasks will you address?
    - **Dataset** - Description of data you are using - the size of the dataset, distribution of classes, any preprocessing you needed to do
    - **Baseline** - Description and implementation of your baseline. You don't need to go into too much detail, but please still include a brief overview.

- **Main Approach** - Propose a model, an algorithm, or a framework for tackling your task. You should describe the model and algorithm in detail and use a concrete example to demonstrate how it works. Don't describe methods in general; describe precisely how they apply to your problem (what are the inputs/outputs, variables, factors, states, etc.)?
- **Evaluation Metric** - Please include what metrics, both qualitative and quantitative, you are using to evaluate the success of your problem.
- **Results & Analysis** - Please include the performance of your baseline as well as the performance of your main approach so far and any experiments that you have run. If your results are creative and can't find a proper baseline, then you can analyze how you get the results you want.
- **Future Work** - This section can be short, but please include some ideas about how you could improve your model if you had more time. This can also include any challenges you're running into and how you might fix them.

## Discussion

TA will open a discussion forum **Final Project** 討論區 on NewE3 of the course, and you can ask questions about the final project in the forum. TA will answer as soon as possible.

## Final Project Report Submission

You have to submit the final project report. In contrast, the initial proposal, mid-term presentation, and final project presentation will be graded on your presentation performance alone, and do not require any file submission.

1. **The submission deadline for the final project report is 12/15 (Mon.) 23:59.**
2. Submit a report with the filename of **Project_Team{TEAM_ID}_report.pdf**. The report should contain all details and your **code link**.
3. **We won't accept any late submissions for the final project.**
4. **We only accept one pdf file**, and the wrong file format or naming format causes -10 points to your score.
5. **Only 1 team member needs to submit the file to NewE3**.
6. (Optional) We highly encourage each team to formulate the project into a paper and submit to a top conference (e.g., ACL2026, SIGIR2026, KDD2026), which will become a strong record on your resume. Feel free to contact TA Yu-Chien Tang and Prof. Yen to discuss more details on the paper submission, and we are extremely welcome to cooperate with you.
7. Please don't hesitate to reach out via email tommytyc.cs13@nycu.edu.tw if you have any questions.

## Topics Introduction

- **Personalization and Recommendations in Search**
    - As users increasingly demand richer and more intuitive search experiences, personalization and recommendation play a central role in modern information retrieval. Search systems are evolving from static query-document matching to dynamic, user-centered services that incorporate context, preferences, behavioral signals, and multimodal cues. The integration of large language models (LLMs), generative recommendation, multimodal representations, and reinforcement learning has transformed personalization in search. These advances enable systems to better capture user intent, mitigate cold-start challenges, support cross-domain learning, and offer conversational and interactive recommendation experiences. Despite these opportunities, key challenges remain: balancing personalization with fairness and privacy, reducing hallucination and bias in generative recommendation, and ensuring scalability and responsiveness at web scale. Addressing these issues will be critical for the next generation of personalized search systems.
    - Topics include but are not limited to:
        - Personalized search and recommendation models
        - Generative recommendation in search
        - Conversational and context-aware recommendation systems
        - Sequential recommendation and user behavior modeling for search
        - Cross-domain personalization and transfer learning
        - Multimodal personalization in search (text, images, video, charts)
        - Knowledge graph–enhanced personalization and recommendation
        - Fairness, privacy, and security challenges in personalized search
        - User interaction and explainability for transparency and trust
        - Real-time and large-scale personalized search system design
        - Adversarial attacks and defenses in personalized search
        - Evaluation metrics: diversity, novelty, coverage, and user satisfaction
        - Human–AI collaboration and interactive design for personalized search
- **Agentic IR (https://sites.google.com/view/ai4ir/aaai-2025)**
    - The field of information retrieval has significantly transformed with the integration of AI technologies. AI agents, especially those leveraging LLMs and vast computational power, have revolutionized information retrieval, processing, and presentation. LLM agents, with tool-call, advanced memory, reasoning, and planning capabilities, can perform complex tasks, engage in coherent conversations, and provide personalized responses. Despite these advancements, challenges such as ensuring relevance and accuracy, mitigating biases, providing real-time responses, and maintaining data security remain.
    - Topics Include but not limited to:

        - Agentic Retrieval System
        - Agentic Conversational Recommendation
        - AI Agent for Personalization
        - AI Agent for Sequential Recommendation
        - AI Agent for Contextual Information Retrieval
        - AI Agents for Cross-lingual and Multimodal Retrieval
        - Retrieval-Augmented Generation Searching System
        - Optimization of AI Agent Retrieval Models
        - Bias Mitigation in AI-driven Information Retrieval
        - Interpretable Generative Retrieval System
        - Adversarial Attacks and Defenses in Agentic Retrieval System
        - Ethical Considerations in Agentic Information Retrieval
        - Human-AI Collaboration in Information Retrieval
        - Evaluation for Agentic Information Retrieval
        - Scalability and Efficiency in AI Agent Retrieval
- **Multimodal Retrieval**
    - As digital content continues to diversify, users increasingly need to search across not only text but also images, videos, audio, tables, and charts. Multimodal retrieval aims to unify representations and understanding across different modalities, enabling search systems to deliver more precise, natural, and interactive access to information. Recent advances in large language models (LLMs), multimodal pre-trained models (e.g., CLIP, BLIP, video-language models), and diffusion-based generative models have accelerated progress in this field. These technologies make it possible to align and embed heterogeneous modalities into shared semantic spaces, supporting tasks such as text-to-image retrieval, speech-to-video retrieval, or text-to-chart retrieval. Nevertheless, multimodal retrieval faces significant challenges: semantic alignment and fusion across modalities, fine-grained retrieval and interpretability, multilingual and cross-cultural understanding, and efficiency at scale. With growing applications in education, healthcare, entertainment, e-commerce, and social platforms, balancing performance, transparency, and fairness is an urgent and important research direction.
    - Topics include but are not limited to:
        - Multimodal retrieval model design (text–image, text–video, audio–visual, tables/charts)
        - Integration of large language models with multimodal retrieval
        - Generative multimodal retrieval
        - Fine-grained retrieval and interpretability (e.g., localizing relevant image/video segments)

        - Multilingual and cross-cultural multimodal retrieval
        - Conversational and interactive multimodal retrieval
        - Real-time and large-scale multimodal retrieval system design
        - Adversarial attacks and robustness in multimodal models
        - Evaluation benchmarks for multimodal retrieval (accuracy, diversity, explainability)
        - Human–AI collaboration in multimodal retrieval applications
- **Domain-Specialized RAG (e.g., Clinical & Legal Focus)**
    - Retrieval-Augmented Generation (RAG) has demonstrated strong potential in open-domain applications, but in specialized domains such as clinical medicine and law, it faces far greater challenges in accuracy, interpretability, and compliance. In these high-stakes fields, tolerance for error is extremely low, and any hallucination or imprecise answer can lead to serious consequences. Thus, tailoring RAG systems to domain-specific requirements is a critical direction for future research. In clinical applications, RAG systems must integrate medical knowledge bases, electronic health records, clinical guidelines, and scholarly literature to provide evidence-based recommendations and explanations, all while safeguarding patient privacy and data security. In legal contexts, RAG must process case law, statutes, contracts, and litigation documents with high precision, ensuring faithful citation, traceability, and support for lawyers and researchers working within complex contexts. Key challenges include building high-quality domain corpora, reducing hallucination while improving citation fidelity, incorporating domain-specific terminology and reasoning, and maintaining both real-time responsiveness and regulatory compliance. As demand grows for AI-assisted systems in healthcare and legal practice, domain-specialized RAG will become increasingly important in both research and real-world applications.
    - Topics include but are not limited to:
        - Domain-specific RAG system design for clinical and legal contexts
        - High-fidelity retrieval and citation generation
        - Integration of domain-specific terminology and knowledge graphs in RAG
        - Structured retrieval and contextual modeling of case law and medical literature
        - Hallucination detection and suppression in domain-specific QA systems
        - Interpretability and traceability in domain-specialized RAG systems
        - Multilingual clinical/legal RAG challenges and opportunities

- Human–AI collaboration: how clinicians and lawyers work with RAG systems
- Adversarial risks and misinformation in professional RAG applications
- Real-world use cases: clinical decision support and legal research assistance
- Benchmarks and evaluation metrics for specialized RAG (accuracy, reliability, explainability)

Reference source: Part of this project requirement is directly inspired by Stanford CS224N.