Final Exam

601.467/667 Introduction to Human Language Technology

Fall 2025
Johns Hopkins University
Co-ordinator: Philipp Koehn

11 December 2025

Complete all questions.

Use additional paper if needed.

Time: 75 minutes.

Name of student: _____

Q1. Question Answering *20 points*

1. Question Answering (QA) is a general problem that can mean different things to different people. Please compare and contrast the three traditions we discussed in class with regard to each of the aspects listed below (one to two sentence per cell) (20 points)

	Classical QA (e.g. IBM Watson)	Machine Reading Comprehension	Large Language Model with RAG
Problem Formulation			
System Architecture			
Usage Scenario			

Q2. Digital Humanities *10 points*

1. What are some *limitations* of traditional scholarship, and corresponding *strengths* of computational methods? (10 points)

Q3. Human-Centered Evaluation *20 points*

You are part of a team building a daily news summarization system that delivers a concise daily brief on political topics. The system must summarize news articles from diverse sources, reflect multiple perspectives (e.g., partisan, regional, ideological), and operate across multiple languages. To evaluate and iterate on the system, you will rely on both human-centered evaluation and automated approaches, including using an LLM as a judge of summary quality.

1. You plan to use an LLM to score the quality of the perspective balance of the system's summaries and to compare candidate summaries. On what dimensions should you evaluate the judge's quality? For each dimension, explain what it is and describe one method to evaluate. (10 points)

2. You are collecting "gold" summaries by recruiting 30 English-speaking-only, computer science students from Johns Hopkins University. They are given political news articles in various languages and use translation software to write summaries. What potential issues could arise from this protocol? (6 points)

3. You plan a between-subjects experiment to compare two news-summarization models (Model A vs. Model B). Each participant is randomly assigned to one model, sees daily political summaries generated by that model, and rates the perceived perspective balance of each summary on a 7-point Likert scale.

   Select all that apply to classify this study:

   A) Empirical

   B) Analytical

   C) Qualitative

   D) Quantitative

   (4 points)

Q4. NLP for Studying Human Behavior *20 points*

1. List at least three limitations that can make LLMs poor proxies of human behavior. (10 points)

2. Name three methods that allow combining imperfect LLM annotations with human annotations for valid inference about human behavior? (6 points)

3. List and define key metrics for evaluating how well these methods work. (4 points)

Q5. Ethical Problems *20 points*

1. How can the principle of explicability enable the principle of justice? (10 points)

2. Explain the concept of informed consent and explain why it is important when collecting data from human subjects. (10 points)

Q6. Computational Social Science *20 points*

You are interested in studying how AI models are impacting education by exploring the perspective of students. You have large dataset of free-form responses to a survey conducted in Spring 2025, 2024, and 2023 among Johns Hopkins undergraduate students. In the survey, students were asked to write 1-2 paragraphs about how they are using AI in their courses and if they believe it helps or hinders their education.

1. What is one way you could analyze this data using a **supervised or classification-based** approach? Be specific in describing exactly what you would analyze and how you would conduct the analysis (2-3 sentences or bullet points). (5 points)

2. What is one way you could analyze this data using an **unsupervised approach**? Be specific in describing exactly what method you would use and what kind results you would hope to uncover (2-3 sentences or bullet points). (5 points)

3. Compare and contrast the supervised/classification and unsupervised approaches you described above. What is 1 key strength and 1 key weakness for each one? (5 points)

4. What are some of the limitations of this study? Include at least two limitations that could undermine the validity of the results. (5 points)