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			

(1) Problem Formulation - Classic QA: Given question and knowledge source (multiple document collection), output answer; MRC: Given question and one document, output answer;

LLM with RAG: Given question (potentially a conversation), a model that ingested a large amount of knowledge, and retrieved documents, output answer.
(2) System Architecture - Classic QA: a pipeline of multiple NLP/IR components; MRC: one neural net trained end-to-end; LLM with RAG - IR system with a large language model.
(3) Usage Scenario - Classic QA: User-facing, supports answering predefined question types; MRC: goal is to test the capability of AI, not to serve an user; LLM with RAG: User-facing, supports a large set of tasks, including question answering.

Q2. Digital Humanities *10 points*

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

   Humans have limited time and attention, and are often invested and biased towards particular conclusions. Computation can separate a method from human error, and explore an empirical hypothesis-space with clearly-stated assumptions and falsifiable predictions.

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)

   1. Reliability and validity (2 points). 2. Clearly defines reliability (consistency/stability of scores) vs. validity (scores measure the intended construct and support correct decisions) (4 points). 3. Reliability evaluation methods and metrics (2 points) Name one concrete reliability check, such as: Test–retest stability across runs/prompts. Internal-consistency reliability within the same dataset. 4. Validity evaluation methods and metrics (2 points) Name one concrete validity check, such as: criterion-based validity. Predictive validity on user outcomes.

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)

   Translation fidelity (3 points): Flags MT errors/low-resource issues affecting accuracy. Cultural/perspective balance (3 points): Mentions missed regional context or US-centric framing.

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)

   A (2 points), D (2 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)

   Possible answers: they do not always align with human judgment, they can encode demographic or other biases, factual inaccuracies, be inconsistent, sensitive to prompts

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

   Prediction-powered inference (2 points), Design-based supervised learning (2 points), Confidence-driven inference (2 points)

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

   Accuracy/coverage (are we retrieving the ground truth estimate) (2 points) and effective sample size (how much value are we getting out of each human annotation/what is the added value of LLM proxies) (2 points)

Q5. Ethical Problems *20 points*

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

   By auditing and understanding what a certain model is doing we can know its biases and evaluate its fairness. In that manner, we can decide to use it or not, or apply measures to counterbalance the biases and avoid unfairness/discrimination, which is an essential part of the justice principle.

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

   Including informed consent on a project that involves data collection entails including a protocol that ensures that the participant understands what is being done and accepts it (researchers must assess if the participant understands what they're signing). The participant must voluntarily consent to the experiment and must have the right to withdraw consent at any time. It is important to protect the privacy and rights of participants. This ensures their data is not used without their consent and that they know how the data is going to be used and the potential risks.

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)

   Label each data point for whether the student thinks the models are helpful or hurtful to education (or any other label that might be reasonable to study in this data). Labeling can be done by hand-annotating a training set and training a small classifier or through LLM-prompting.

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)

   Use topic modeling (LDA, TopicGPT, CTM) to analyze broad trends in the data.

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)

   Supervised: easier to control, can target specific research questions, but can be time-consuming to set-up and validate. Unsupervised: able to run with minimal initial manual labeling, can study exploratory questions without knowing specific research questions in advance, can be difficult to control and can reach the wrong conclusions if patterns are misinterpreted

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)

   Examples: students may not have been fully honest about AI usage and its impact, JHU students may not be representative of other universities, inaccurate/poorly validated models could lead to incorrect conclusions