

University of Central Missouri
Department of Computer Science & Cybersecurity

CS5760 Natural Language Processing

Fall 2025

Homework 5.

Student name: DUGGINENI SESHA RAO

Submission Requirements:

- Once finished your assignment push your source code to your repo (GitHub) and explain the work through the ReadMe file properly. Make sure you add your student info in the ReadMe file.
- Submit your GitHub link on the Bright Space.
- Comment your code appropriately ***IMPORTANT***.
- Any submission after provided deadline is considered as a late submission.

Part I. Short Answer

Q1. Ethical Foundations

- a) Explain why *ethics* is not the same as *laws* or *feelings*.

Ethics is not the same as laws or feelings because:

- **Laws** are government rules and may allow unethical things or ban morally right ones.
- **Feelings** are personal reactions and can be biased or uninformed, but ethics is based on reasoning about right and wrong.

- b) Briefly describe two classical ethical theories (e.g., utilitarianism and deontology) and how they would handle an AI decision scenario.

classical ethical theories:

- **Utilitarianism:** Focuses on outcomes. A decision is ethical if it maximizes overall happiness.

AI example: Approve an automated medical treatment if it saves the greatest number of lives.

- **Deontology:** Focuses on rules and duties. A decision is ethical if it follows moral principles.

AI example: Reject an action that violates privacy, even if it benefits many people

- c) Why do philosophers argue that no single ethical theory clearly “wins” in all contexts?

No single theory always “wins” because real-world situations involve competing values (e.g., fairness vs. maximizing benefit), and different theories prioritize different moral goals.

Q2. Types of AI Harms

- a) Define allocational harm and representational harm in AI systems.

Allocational harm: When AI affects access to opportunities or resources (e.g., jobs, loans).

Representational harm: When AI spreads stereotypes or treats groups disrespectfully.

- b) Provide an example of each from real-world applications (e.g., translation, hiring, or facial recognition).

- Allocational: Hiring algorithm rejecting qualified female candidates more often than men.

- Representational: Translation model using biased gender roles like “doctor → he” / “nurse → she”.
 - c) Why is representational harm often harder to measure than allocational harm?
- Representational harm is harder to measure because it relates to dignity, stereotypes, and cultural impact—not just numbers and outcomes.

Q3. Sources of Dataset Bias

- a) List three reasons why bias arises during data collection or annotation in AI datasets.

Sampling bias: Data collected from only certain locations or demographics.

Historical bias: Existing societal inequalities reflected in data.

Annotation bias: Human labelers carry personal judgments or cultural assumptions.

- b) What kinds of data or groups tend to be under-represented in large language datasets?

Groups often under-represented:

- Low-resource languages
- Minorities and marginalized communities
- People with disabilities and older adults

- c) How can bias amplification occur even after initial data preprocessing?

Bias amplification can happen during model training because learning algorithms may exaggerate existing patterns to improve accuracy (e.g., reinforcing stereotypes).

Q4. Safety, Security, and Privacy

- a) Define data poisoning and describe how it can manipulate a model’s predictions.

Data poisoning:

Attackers intentionally insert corrupted data into a dataset so the model learns wrong behavior, causing manipulated outputs (e.g., misclassifying a stop sign as a speed limit)

- b) What are the ethical implications of model memorization (e.g., GPT-2 reproducing private or copyrighted text)?

Model memorization risks:

- May leak **private data**, like personal information from training set
- May reproduce **copyrighted or proprietary text**, raising legal and ethical concerns

c) How does model stealing threaten privacy and intellectual property in AI research?

If attackers copy a model's functionality through repeated queries, they can steal **intellectual property** and potentially expose **sensitive training information** that threatens privacy.

Q5. Harm Mitigation and Best Practices

a) Describe one technical and one non-technical approach for mitigating bias or harm in AI systems.

Technical method: Fairness-aware training (e.g., re-weighting under-represented groups).

Non-technical method: Inclusive stakeholder review or impact assessments.

b) What is the goal of tools such as Model Cards and Bias Audits?

Purpose of Model Cards & Bias Audits:

To document how a model was trained, its limitations, and fairness checks—ensuring transparency and accountability.

c) Explain what is meant by *expanding the ethical circle* and why it is important for inclusive AI development.

Expanding the ethical circle:

Including more people and groups (especially marginalized ones) in ethical consideration.

This is important to ensure AI benefits everyone, not only those already privileged.