# Large Language Model project cards

Academic Year 2024-2025

# Project assignment

- Teams of 2-3 people
- Select five project proposals (at least one for category) that you would like to do
- We will assign you - if possible - one of the projects you have chosen

- Deadline for team and proposal selection: November 30
- Assignment of projects and project start: December 1

- https://docs.google.com/spreadsheets/d/1SRntBPtHONhRIPhkoRlobD4lQ9GvpfPKuRbyAw4C8Hw/edit?usp=sharing

# Project evaluation

- Deadline for project hand-in: before the beginning of next academic year (September, 2025)

- Project points: 18/30

- The project score will be valid for the 2024-2025 academic year and registered when you first pass the written exam

# Project delivery

- You will receive a document with a template that you have to fill in with your methodology and result

- To deliver the project, you have to submit:

  1. The document describing your project work, as a technical report created with Overleaf (or an equivalent desktop or web LaTeX client) https://www.overleaf.com/latex/templates/latex-template-for-technical-report/qtznkrpkjybm

  2. The link to a GitHub project

# Examples of project delivery

You need to fill a pre-defined technical report (6-8 pages) with some mandatory sections.

The technical report will already contain some useful information for methodology and reporting.

For the Experiment projects, refer to the documentation of SemEval:
https://semeval.github.io/SemEval2025/tasks

Example for Application project:
https://docs.google.com/document/d/1dZg7uuTznZPjZl48N7NfCtHy23uChQnm1q_vp6Yf21Y/edit?usp=sharing

Example for Review project:

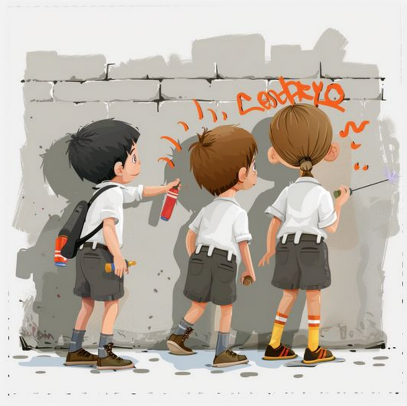
https://docs.google.com/document/d/1fSjhPrmc1UwBC-Z7y0biVleF4eIVt7E5vi_TiqNBBkc/edit?usp=sharing

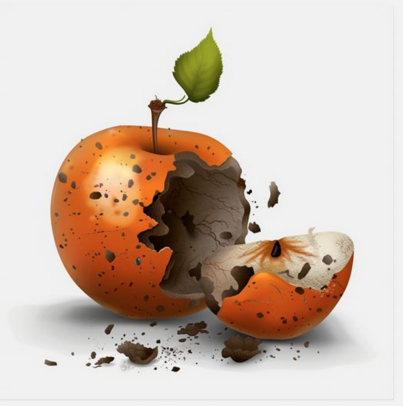
# Categories of projects

- Experiment: technical projects in which you will analyze and compare internal properties and technical characteristics of LLM models.
- Application: projects in which you will analyze the effectiveness of the application of LLMs in various Software Engineering tasks.
- Review: *systematic* analysis of the literature about LLMs, to perform a detailed and critical assessment of some crucial aspects of the state of the practice.

# E1: ADMIRE

Advancing Multimodal Idiomaticity Representat

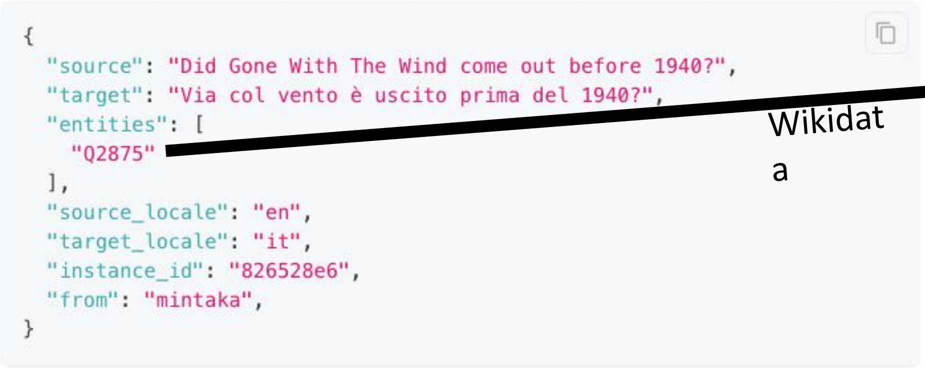




- Which of these images best represents the meaning of the phrase ***bad apple*** in the following sentence?
  - "We have to recognize that this is not the occasional *bad apple* but a structural, sector-wide problem"
  - "However, if ethylene happens to be around (say from a *bad apple*), these fruits do ripen more quickly."
- Computational language models struggle with figurative expressions
- **Subtask A**: Static Images (5 images + context sentence □ rank images according to relevance)
- **Subtask B**: Next Image Prediction (2 images + context sentence □ choose next best image)

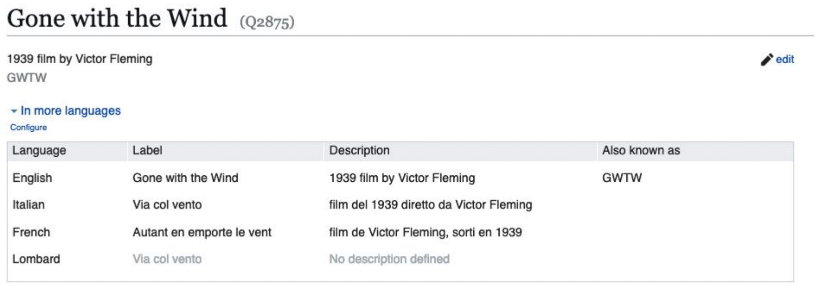
# E2: EA-MT
# Entity-Aware Machine Translation

- The task is to translate a given input sentence from the source language (English) to the target language, where the input sentence contains *named entities that may be challenging* for machine translation systems to handle
  - **English Sentence**: "I bought a new book called *'The Catcher in the Rye*'."
  - **Italian Sentence**: "Ho comprato un nuovo libro chiamato *'Il Giovane Holden*'."



```
{
  "source": "Did Gone With The Wind come out before 1940?",
  "target": "Via col vento è uscito prima del 1940?",
  "entities": [
    "Q2875"
  ],
  "source_locale": "en",
  "target_locale": "it",
  "instance_id": "826528e6",
  "from": "mintaka",
}
```

Wikidata



Gone with the Wind (Q2875)

1939 film by Victor Fleming

GWTW

| Language | Label | Description | Also known as |
|---|---|---|---|
| English | Gone with the Wind | 1939 film by Victor Fleming | GWTW |
| Italian | Via col vento | film del 1939 diretto da Victor Fleming | |
| French | Autant en emporte le vent | film de Victor Fleming, sorti en 1939 | |
| Lombard | Via col vento | No description defined | |

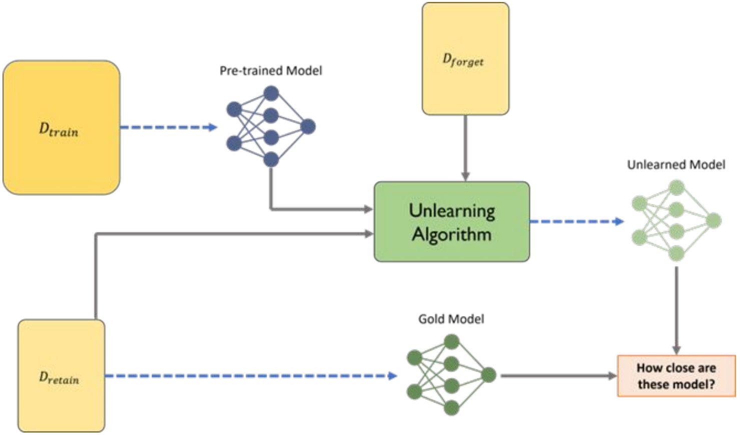
# E3: Mu-SHROOM

Multilingual Shared-task on Hallucinations and Related Observable **Overgeneration Mistakes**

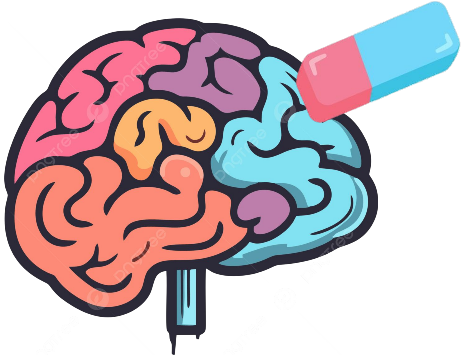
- The objective is to detect hallucination spans in the outputs of instruction-tuned LLMs in a multilingual context:
  - Example:
    - Prompt: "Over which state did Xiong Ai rule?"
    - Model output: "Xiong Ai was a Chinese warlord who lived during the Warring States period (475-221 BC). He ruled over the state of Xiongnu, which is modern-day Mongolia rather than a Chinese state. The Xiongnu were an influential nomadic confederation that posed a significant threat to China during this time. Therefore, Xiong Ai's rule was not over a Chinese state but rather over the Xiongnu people"

# E4: Unlearning sensitive content from Large Language Models

- The challenge's objective is to **remove the influence of selected information ($D_{forget}$) from a LLM's training.**
  - Some short false biographies were included in the training of an LLM. Your goal will be to make the model forget this information so that it behaves in the same way as a model trained only with the remaining part of the training set ($D_{retain}$)
  - AMAZON organizes the challenge, and a 1B parameters model is available!



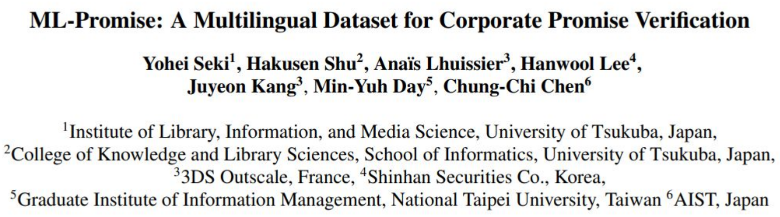
Example of Machine Unlearning Pipeline



# E5: LLM-based Subject Tagging for the TIB Technical Library's Open-Access Catalog

- Goal: Tag technical records with a given taxonomy
- Two subtasks:
  1. **LLM-based solution for subject tagging** of technical records from Leibniz University's Technical Library (TIBKAT).
     - Requires bilingual language modeling (German and English)
  2. **Align Subject Tagging** to the TIBKAT collection.
     - Align subject tagging capability of their systems with the annotations provided in TIBKAT.

# E6: PromiseEval: Multinational, Multilingual, Multi-Industry Promise Verification



**ML-Promise: A Multilingual Dataset for Corporate Promise Verification**

**Yohei Seki[1], Hakusen Shu[2], Anaïs Lhuissier[3], Hanwool Lee[4],**
**Juyeon Kang[3], Min-Yuh Day[5], Chung-Chi Chen[6]**

[1]Institute of Library, Information, and Media Science, University of Tsukuba, Japan,
[2]College of Knowledge and Library Sciences, School of Informatics, University of Tsukuba, Japan,
[3]3DS Outscale, France, [4]Shinhan Securities Co., Korea,
[5]Graduate Institute of Information Management, National Taipei University, Taiwan [6]AIST, Japan

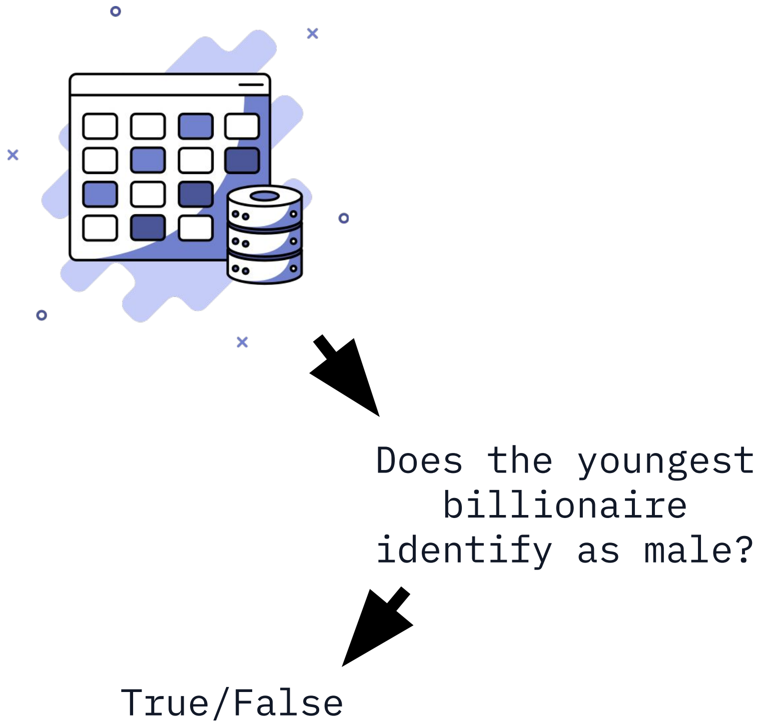
- The objective of this challenge is to assess the idea of **«PROMISE VERIFICATION»**
  - “Recognizing the critical role of transparency and accountability in today's society. (…) In the evolving landscape of Environmental, Social, and Governance (ESG) criteria, the ability to accurately assess a company's commitment and adherence to its ESG promises has become paramount.
  - Labels evaluated in the dataset, composed of ESG reports from different companies, markets, and languages:
    - **Promise Identification**: This is a boolean label (Yes/No) based on whether a promise exists.
    - **Supporting Evidence**: This is a boolean label (Yes/No) based on whether supporting evidence exists.
    - **Clarity of the Promise-Evidence Pair**: Three labels (Clear/Not Clear/Misleading), which represent the clarity of the given evidence with the promise.
    - **Timing for Verification**: Following the MSCI guidelines, we set timing labels (within 2 years/2-5 years/longer than 5 years/other) to indicate when readers/investors should return to verify the promise. Here, "other" denotes the promise has already been verified or doesn't have a specific timing to verify it.

# E7: Multilingual and Cross-lingual Fact-Checked Claim Retrieval

- The objective of this task is to develop a system to **retrieve relevant fact-checked claims** for given social media posts across multiple languages.
  - In this task, you are given *social media posts* (SMP), and a bunch of *fact-checks* (FC). The goal is to find the most relevant fact-checks for each social media post.

# E8: DataBench, Question-Answering over Tabular Data

- The objective of this task consists of **Question Answering over real-world Tabular Data** from different domains
  - Participants will be provided with a tabular dataset (of any size) and a question over it. The question should be answered using the data from the dataset only.

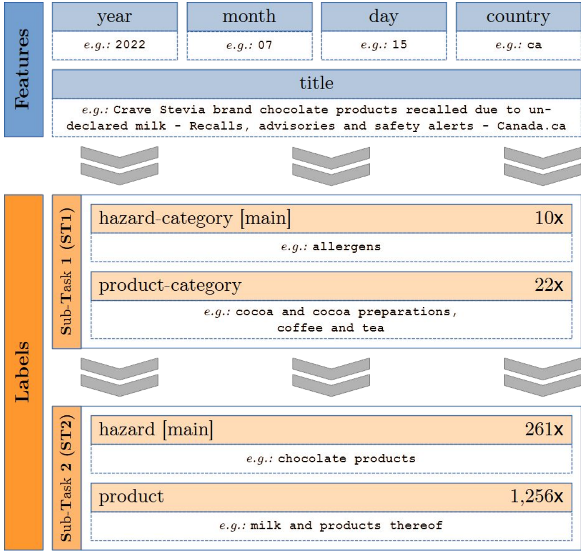


Does the youngest billionaire identify as male?

True/False

# E9: The Food Hazard Detection Challenge

- The objective of this task consists of developing an **explainable classification system for titles of food-incident reports** collected from the web
  - Participants will base their analysis on either the *"title"* or the *"text"* feature (indicating which one they used). The task is to predict the labels *"product-category"* and *"hazard-category"* and the vectors *"product"* and *"hazard"*.
    - **22 categories** (e.g., "meat, egg and dairy products," "cereals and bakery products," "fruits and vegetables")
    - **128** possible **hazard** values (e.g., "salmonella," "listeria monocytogenes," "milk and products thereof"), sorted into **10 hazard-category** values.



Features

year
*e.g.:* 2022

month
*e.g.:* 07

day
*e.g.:* 15

country
*e.g.:* ca

title
*e.g.:* Crave Stevia brand chocolate products recalled due to undeclared milk - Recalls, advisories and safety alerts - Canada.ca

Labels

Sub-Task 1 (ST1)

hazard-category [main] 10x
*e.g.:* allergens

product-category 22x
*e.g.:* cocoa and cocoa preparations, coffee and tea

Sub-Task 2 (ST2)

hazard [main] 261x
*e.g.:* chocolate products

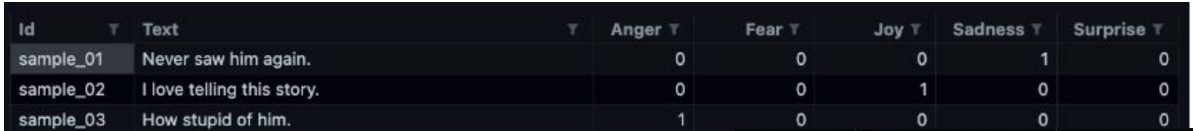
product 1,256x
*e.g.:* milk and products thereof

# E10: Multilingual Characterization and Extraction of Narratives from Online News

- This task challenges participants to **analyze news articles and automatically identify narratives, classify them, and determine the roles played by relevant entities**. The task is multilingual and covers five languages
  - Subtask 1: **Entity Framing** – Classify the roles of named entities within news articles.
  - Subtask 2: **Narrative Classification** – Classify each article based on all the (sub)narratives given a specific domain.
  - Subtask 3: **Narrative Extraction** – Generate short textual explanations for dominant narratives in the articles.

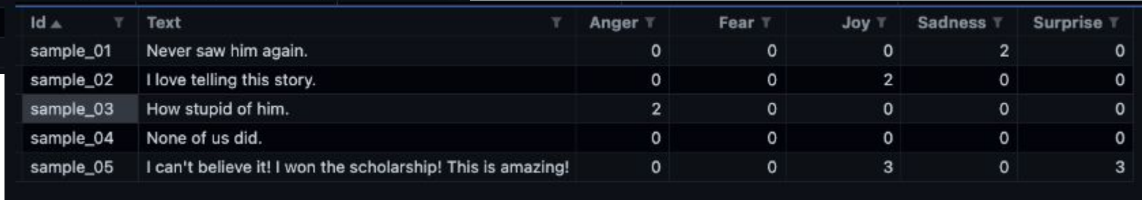
# E11: Bridging the Gap in Text-Based Emotion Detection

- This task challenges participants to classify which one could be the **perceived emotion** from a text:
  - The objective is to determine what **emotion most people will think the speaker may be feeling given a sentence** or a short text snippet uttered by the speaker.
  - The task is **not** about emotion evoked in the speaker or even the genuine emotion of the speaker!



| Id | Text | Anger | Fear | Joy | Sadness | Surprise |
|---|---|---|---|---|---|---|
| sample_01 | Never saw him again. | 0 | 0 | 0 | 1 | 0 |
| sample_02 | I love telling this story. | 0 | 0 | 1 | 0 | 0 |
| sample_03 | How stupid of him. | 1 | 0 | 0 | 0 | 0 |
| sample_04 | None of us did. | 0 | | | | |
| sample_05 | I can't believe it! I won the scholarship! This is amazing! | 0 | | | | |

**Track A: Multi-label Emotion Detection**



| Id | Text | Anger | Fear | Joy | Sadness | Surprise |
|---|---|---|---|---|---|---|
| sample_01 | Never saw him again. | 0 | 0 | 0 | 2 | 0 |
| sample_02 | I love telling this story. | 0 | 0 | 2 | 0 | 0 |
| sample_03 | How stupid of him. | 2 | 0 | 0 | 0 | 0 |
| sample_04 | None of us did. | 0 | 0 | 0 | 0 | 0 |
| sample_05 | I can't believe it! I won the scholarship! This is amazing! | 0 | 0 | 3 | 0 | 3 |

**Track B: Emotion Intensity**

# SemEval 2025

A thorough description of all tasks is available on the SemEval 2025 webpage:

https://semeval.github.io/SemEval2025/tasks

# A1: Goal oriented API alignment

**Goal-Oriented Requirements Engineering (GORE)**

The final objective of GORE is the identification of all goals of the system, defined as Objectives that the system under consideration should achieve. Goals can be formulated at different levels of abstraction, ranging from high-level strategic concerns to low-level technical ones. The main pillars of the GORE technique are the following:

- Goal Modeling: modeling goals according to intrinsic features (e.g., goal type and goal attributes) and links to other goals or other elements of a requirements model (e.g., actors of the system).

- Goal Specification: precise specification of goals to support requirements elaboration.

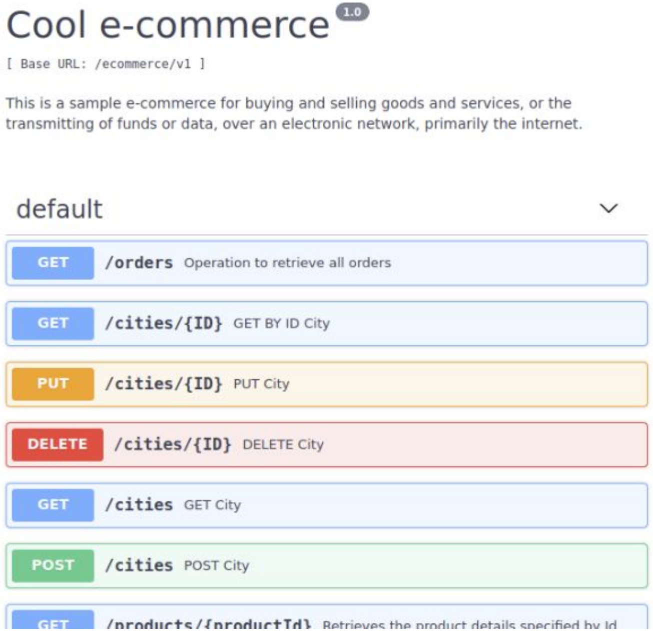
- Goal Reasoning: elaboration of the goal by verifying that they correspond with the requirements of the system; validating the goals by identifying scenarios covered by them, and operationalizes the goals.

**Research Questions:**

*RQ1: What is the effectiveness of LLM agents in modeling high-level goals?*

*RQ2: What is the effectiveness of LLM agents in decomposing high-level goals to low-level goals?*

*RQ3: What is the effectiveness of LLM agents in mapping low-level goals to API endpoints?*



Cool e-commerce 1.0

[ Base URL: /ecommerce/v1 ]

This is a sample e-commerce for buying and selling goods and services, or the transmitting of funds or data, over an electronic network, primarily the internet.

default

GET /orders Operation to retrieve all orders

GET /cities/{ID} GET BY ID City

PUT /cities/{ID} PUT City

DELETE /cities/{ID} DELETE City

GET /cities GET City

POST /cities POST City

GET /products/{productId} Retrieves the product details specified by Id

# A2: Intelligent code quality assessment and refactoring

In the field of software development, recent advancements in artificial intelligence have enabled the application of Large Language Models (LLMs) to code quality analysis. As a framework for defining and evaluating code quality, the focus of the project is on the potential of LLMs to not only assess existing code in GitHub repositories but to also provide actionable recommendations for improvement. Code quality in this context encompasses various aspects, from structural clarity and adherence to programming best practices, to modularity and maintainability over time.

The primary objectives of this approach include the following:

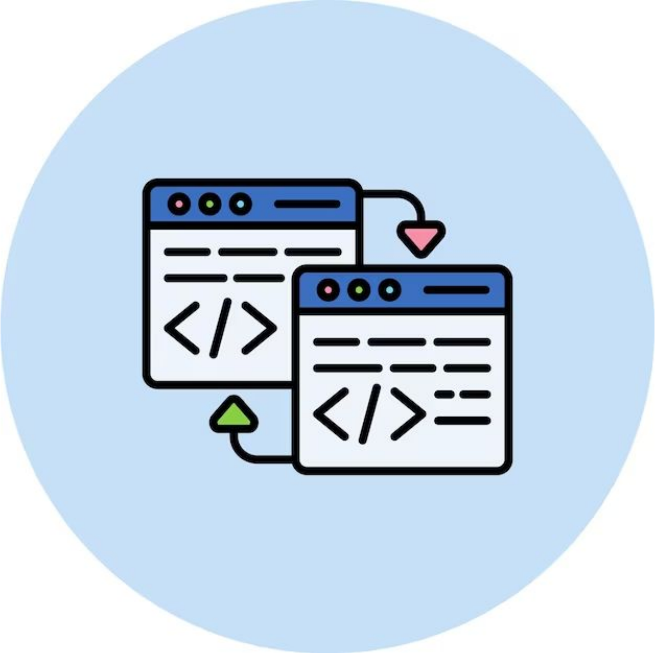
· Code Quality Assessment: Evaluating code according to intrinsic quality metrics, such as readability, maintainability, and adherence to best practices.

· Refactoring Recommendation: Identifying areas of the code that may benefit from refactoring, suggesting targeted improvements for increased modularity, performance, or simplicity.

**Research Questions:**

*RQ1: How effectively can LLMs identify code quality issues in existing codebases?*

*RQ2: To what extent can LLMs provide actionable refactoring recommendations for enhancing code structure and quality?*



# A3: Context-Aware GUI Testing for Mobile Applications

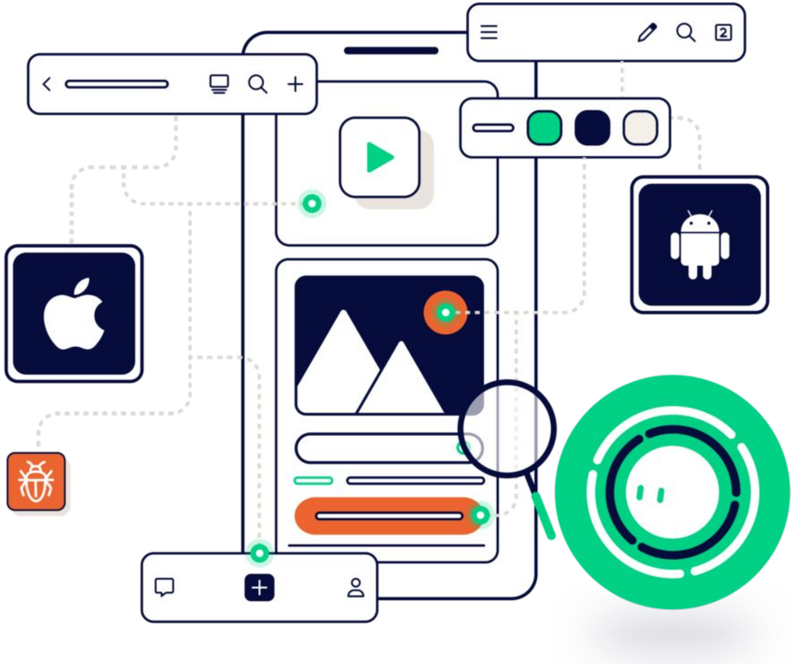
In recent years, testing Graphical User Interfaces (GUIs) for mobile applications has faced increasing challenges due to the diversity and complexity of mobile app interactions. Traditional automated testing tools often lack the adaptability to handle context-specific user interactions, dynamic content, and the wide variety of screen sizes and resolutions in mobile devices. In this context, the application of Large Language Models (LLMs) has emerged as a promising approach for enhancing GUI testing by making it more context-aware and adaptable to real-world scenarios.

LLMs, trained on extensive datasets of language and user interactions, offer the potential to understand user intent and app behavior within different contexts. This ability allows LLMs to simulate realistic user interactions and predict potential edge cases that are difficult to capture with rule-based automation alone. For example, LLMs can generate test cases that respond to contextual cues, such as location, app permissions, or user settings, and adapt these tests as app states change dynamically.

**Research Questions:**

*RQ1: How effectively can LLMs generate context-aware test cases for mobile GUI testing?*

*RQ2: What coverage can be obtained by generating test cases with LLMs?*



# A4: Adaptive GUI Test Evolution and Oracle Maintenance

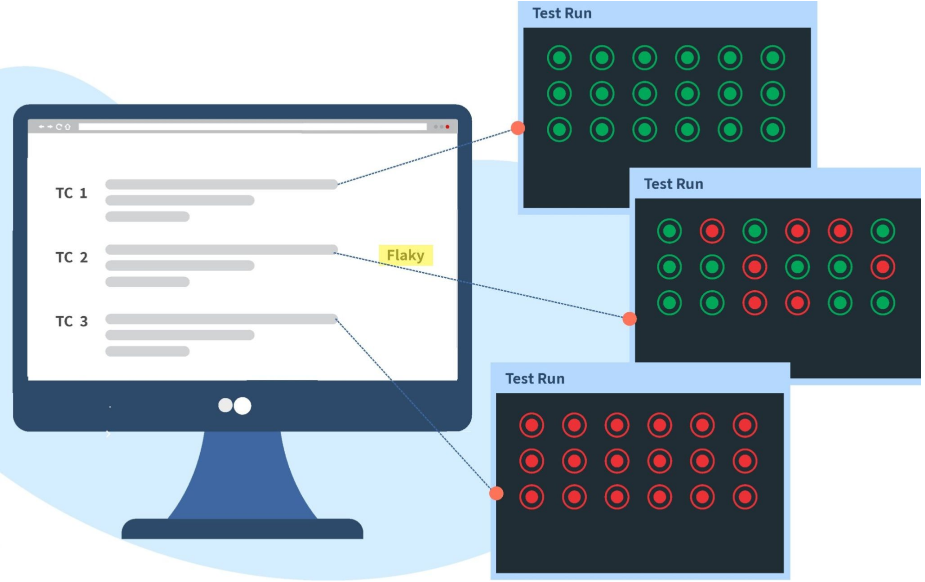
Important challenges in GUI testing pertains the area of test evolution and maintenance. As applications evolve, test suites must keep up with frequent GUI changes, which can lead to the "oracle problem": determining the expected outcomes or "oracles" of test cases in the face of these updates. This issue is often compounded by ambiguities in user interactions and dynamically changing content, making it difficult to maintain accurate and meaningful test assertions. In this context, Large Language Models (LLMs) present a promising solution for enhancing GUI test evolution and repair by providing context-aware, adaptable insights into test oracles.

LLMs, with their advanced language understanding and reasoning capabilities, offer the potential to address the oracle problem by generating or refining test oracles based on app context, expected user interactions, and historical data. This enables more accurate detection of changes in app behavior and helps determine whether the observed changes align with intended functionality or represent defects. For instance, LLMs can analyze changes in GUI layout or text and suggest updates to assertions, adapting tests to evolving requirements and preserving test validity over time.

**Research Questions:**

*RQ1: How effectively can LLMs generate context-aware test cases for mobile GUI testing?*

*RQ2: What coverage can be obtained by generating test cases with LLMs?*



# A5: Code Review and Project Workflow Analysis for Git Data

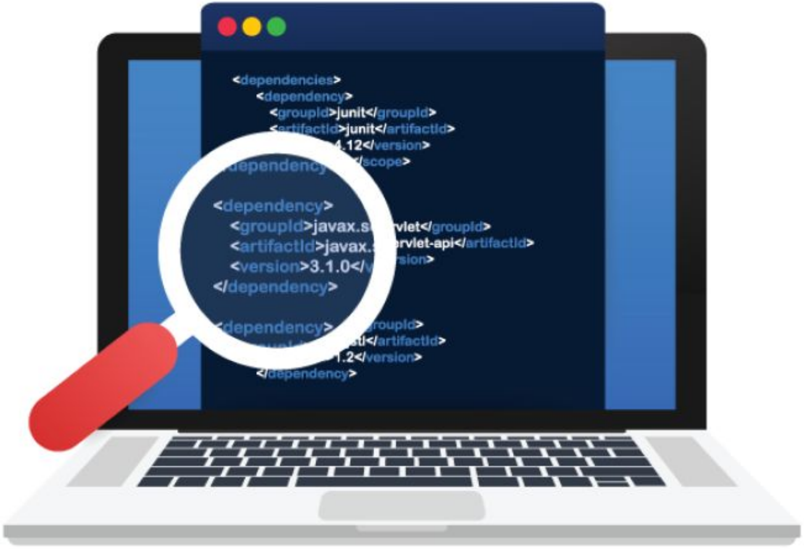
In software development, tracking and understanding the evolution of a complex codebase is critical for maintaining code quality, guiding future development, and supporting informed decision-making. Traditional code review processes, however, can struggle with the vast amount of changes, especially in complex projects with intricate histories, diverse team contributions, and numerous interdependencies. Large Language Models (LLMs) offer a promising approach to address this challenge by leveraging their capacity to analyze and interpret git commit messages, diffs, and tree structures, generating insights into the project's evolution and enhancing the code review process.

With their extensive language and pattern recognition capabilities, LLMs can process and analyze git data to summarize, categorize, and contextualize code changes over time. This provides development teams with structured, meaningful insights into commit intentions, the rationale behind changes, and the impact on the overall project. For instance, LLMs can identify and group commits related to specific features, highlight areas of frequent modification, or detect patterns in code refactoring and bug fixes, offering developers a clear view of how a project has evolved.

**Research Questions:**

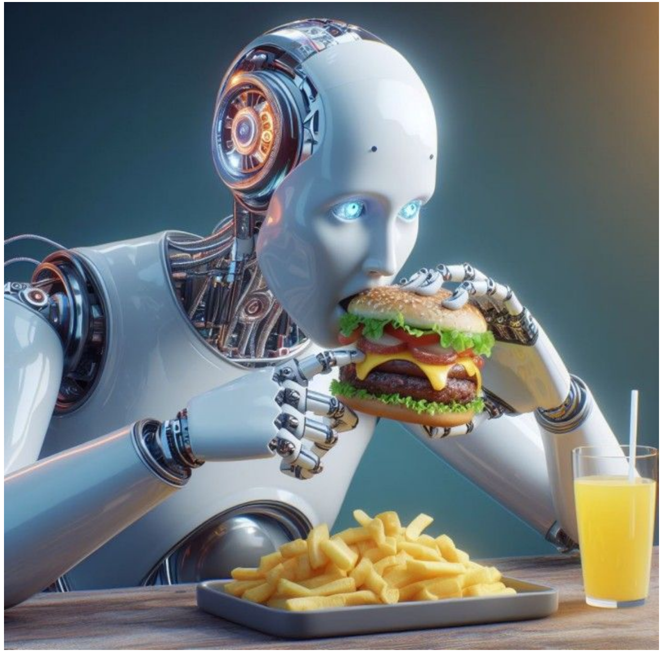
*RQ1: How effectively can LLMs summarize and interpret git commit messages to convey the intent behind code changes?*

*RQ2: How can LLMs facilitate better communication among team members by generating context-aware reports on project evolution?*



# A6: Well-being suggestions

Large Language Models (LLMs) have significant potential to enhance well-being by providing personalized suggestions grounded in medical data. These models can process vast amounts of health-related information, including electronic health records (EHRs), medical research, and patient-provided data, to deliver tailored recommendations. By analyzing patterns in an individual's medical history, lifestyle, and genetic predispositions, LLMs can suggest preventative measures, dietary adjustments, or exercise routines that align with evidence-based practices. For example, an LLM might recommend specific nutritional changes for a patient with a history of diabetes or suggest stress-management techniques for someone prone to anxiety. The adaptability and scalability of these models make them valuable for both patients and healthcare providers, fostering proactive health management.



**Research Questions:**

*RQ1: How effectively can LLMs summarize and interpret collections of medical data to provide well-being suggestions?*

*RQ1: How effectively can LLMs summarize and interpret collections of nutritional data to provide well-being suggestions?*

# R1: LLMs in Healthcare and Medicine

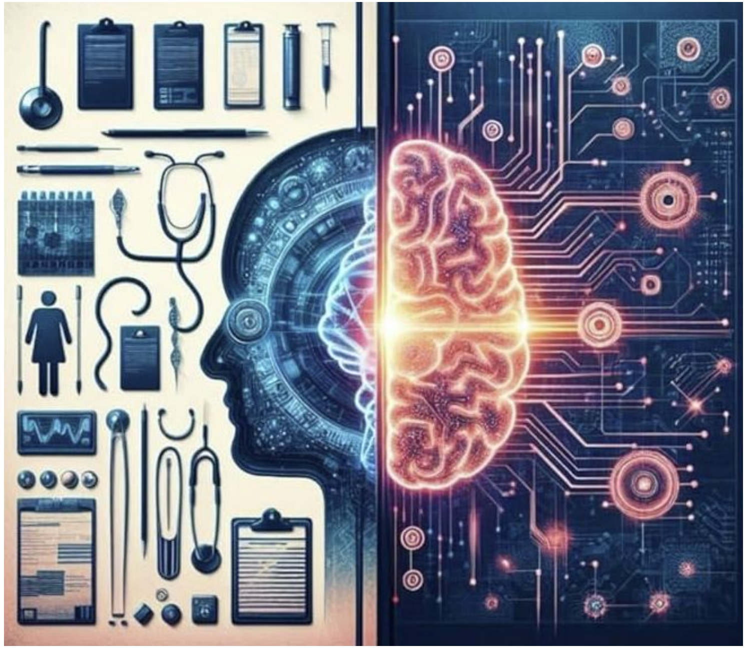
Large language models (LLMs) are revolutionizing healthcare by aiding in diagnostics, medical research, patient support, and personalized medicine. They analyze extensive datasets, such as medical records and scientific literature, to assist clinicians in diagnosing conditions, recommending treatments, and providing preliminary medical advice through chatbots and virtual assistants, improving accessibility and alleviating the burden on medical professionals. However, the use of LLMs in healthcare raises significant ethical concerns, including data privacy issues, the accuracy and reliability of AI-generated medical advice, and the potential for biased outcomes that may exacerbate health disparities. Additionally, the lack of transparency in decision-making processes complicates trust and accountability, highlighting the need for careful oversight to ensure the safe and equitable integration of LLMs in healthcare.

**Research Questions:**

*RQ1: How are LLMs being applied in healthcare, and what are the main benefits?*

*RQ2: What risks do LLMs pose in terms of data privacy and accuracy in medical contexts?*

*RQ3: How do existing regulations address the use of LLMs in healthcare?*



# R2: LLMs in Education and Knowledge Dissemination

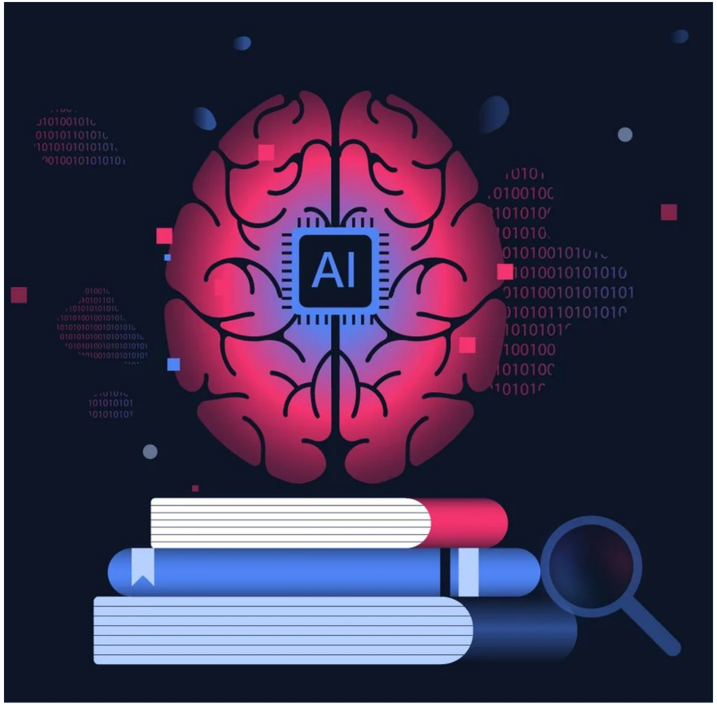
Large language models (LLMs) are revolutionizing education and knowledge dissemination by providing personalized learning support, instant explanations, and tools for generating study materials and assisting with writing. They help automate administrative tasks, allowing educators to focus more on teaching and mentoring, while also assisting in research by summarizing complex texts and simplifying technical concepts. However, the use of LLMs in education raises ethical concerns, including the potential spread of misinformation, over-reliance on AI-generated content, and challenges related to intellectual property and academic integrity. The lack of transparency in LLM decision-making further complicates the verification of information, making it essential to balance the benefits of LLMs with these ethical issues to ensure they enhance, rather than compromise, educational quality.

**Research Questions:**

*RQ1. What roles do LLMs play in enhancing or transforming educational processes?*

*RQ2. How do ethical concerns (e.g., dependency, misinformation) impact the use of LLMs in education?*

*RQ3. How are educational institutions addressing biases and inaccuracies from LLM outputs?*



# R3: LLMs in Creative Industries and Content Generation

Large language models (LLMs) are transforming creative industries by enabling faster and more diverse content generation in fields like marketing, journalism, publishing, and entertainment. They assist with tasks such as drafting written content, brainstorming ideas, and even generating scripts, allowing creators to experiment with new formats and expand their output. However, LLM adoption raises ethical concerns, including issues of originality, intellectual property, and the potential reduction of job opportunities for human creators. Furthermore, LLM-generated content may mislead audiences about its authenticity, and biases in training data can perpetuate harmful stereotypes. Balancing the efficiency of LLMs with these ethical challenges is crucial for their sustainable use in creative industries.

**Research Questions:**

*RQ1: How are LLMs shaping content generation and what benefits or challenges arise?*

*RQ2: What are the ethical implications for originality and copyright in creative industries?*

*RQ3: How does LLM usage affect the job landscape in content creation roles?*

# R4: The Environmental and Social Impacts of LLMs

The development and deployment of large language models (LLMs) have significant environmental and social impacts. Training LLMs requires massive computational power, consuming substantial energy and contributing to a large carbon footprint, raising concerns about the sustainability of AI technologies. Additionally, the hardware for LLMs depends on rare earth minerals, further straining global supply chains and contributing to environmental degradation. Socially, LLMs may cause job displacement in sectors like customer service and content creation, exacerbate digital inequity by limiting access to powerful tools, and reinforce biases from training data, disproportionately affecting marginalized groups. Addressing these challenges is crucial to ensure LLM development aligns with sustainability, fairness, and social responsibility.

**Research Questions:**

*RQ1: What is the environmental cost of developing and deploying large LLMs, and how can it be mitigated?*

*RQ2: How do LLMs affect employment and what societal changes might they drive?*

*RQ3: What steps are being taken to ensure equitable access to LLM technology?*

# R5: Diversity, Bias, and Racial Stereotypes in LLMs

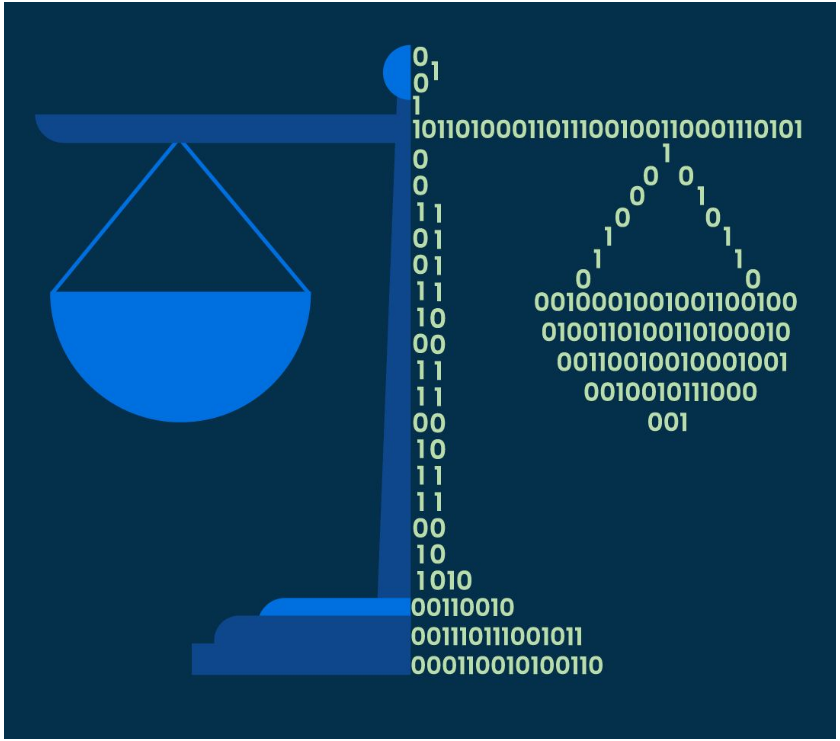
Large language models (LLMs) often reflect and amplify biases present in the datasets they are trained on, which can result in the perpetuation of racial stereotypes and other forms of discrimination. Trained on vast amounts of data from the internet, LLMs can internalize societal prejudices, generating biased outputs that associate certain traits, professions, or behaviors with specific races, genders, or ethnicities. These biases can cause harm when LLMs are used in applications such as chatbots or content generators, unintentionally reinforcing harmful stereotypes and diminishing inclusivity. To address these issues, strategies such as creating more diverse datasets, applying fairness algorithms, and implementing debiasing techniques are being explored. However, these methods are not always fully effective and may reduce model performance. Furthermore, the lack of transparency in LLM training processes complicates efforts to hold developers accountable for biased outputs. These challenges highlight the need for ethical guidelines, regulatory frameworks, and increased oversight to ensure that LLMs promote diversity, equity, and social responsibility in their deployment.

**Research Questions:**

*RQ1: How do LLMs inherit and reinforce racial and cultural biases from their training data?*

*RQ2: What are the specific challenges in detecting and quantifying bias and harmful stereotypes in LLMs?*

*RQ3: What ethical frameworks or guidelines are being proposed or implemented to handle diversity and reduce harmful biases in LLMs?*



# R6: Privacy and Security Challenges in LLM Deployment

The deployment of large language models (LLMs) in real-world applications brings significant privacy and security challenges that need careful consideration. As LLMs are trained on vast datasets, they often require access to sensitive data, such as personal information, medical records, or financial details, depending on the application. This raises concerns about data privacy, especially when LLMs are used in sectors like healthcare, finance, and customer service, where data security is paramount. The risk of data leakage, either through model outputs or adversarial attacks, is a significant concern, as LLMs can inadvertently generate or reveal private information from their training data. Furthermore, the models themselves might be susceptible to reverse-engineering or "model inversion," where attackers can extract sensitive information from a trained model, potentially compromising the privacy of individuals whose data was included in the model's training.

**Research Questions:**

*RQ1: What are the primary privacy risks associated with the deployment of large language models (LLMs) across different industries, and how can these risks be mitigated?*

*RQ2: How do adversarial attacks and model inversion techniques compromise the security of LLMs, and what methods are being developed to enhance their robustness against such threats?*

*RQ3: What are the ethical implications of data leakage and the potential exposure of private information in LLM-generated outputs, and how can transparency and accountability be improved in LLM deployment?*