[redacted] [redacted]

Summative Assignment Part 1 & Part 2

Deadline for summative assignment Part I submission: 14 November 2024, 14:00 GMT, Ultra
Deadline for summative assignment Part II submission: 14 January 2025, 14:00 GMT, Ultra

1. LEARNING OBJECTIVES

Students will experiment with **Think-Aloud** techniques [1, 2, 3, 5] to collect **thought data** and synchronise it with **action data**. Students will also deploy an AI-infused **emotion recognition tool (ERT)** to capture affective responses when interacting with an **LLM-powered application (LLMA)** and evaluate **trust** in the LLMA**.** The main learning outcome is to develop a data-driven understanding of how LLMAs and ERTs are integrated into people's everyday life and their impacts by reflecting on personal practice and experience. The data collected will allow students:

- to evaluate how thought data (with or without action data) can improve AI performance, analysing the potential of the **Thought Cloning** [4] approach.
- to visualise how emotional states change over time and influence human-AI interaction.
- to explore how emotion, thought, action, and trust interrelate during interactions.

Note: No AI model development is required. The project will focus on data collection, pattern analysis, and critical thinking.

2. WORKFLOW

A. PLAN

1. Tool selection:

- Select and use existing **LLMAs** that you deem fit to support your academic as well as non-academic activities, and an **ERT** that can automatically detect emotions and display results. Emotion analysis can be performed in a **mono-modal** (voice only) or **multimodal** (voice + facial expression) mode. The latter can provide richer data and better insights into the topics investigated.
- Use **the same ERT** for all LLM-based activities. Here are some examples of LLMAs and ERTs (Appendix A in Ultra; explore more on your own).

2. User Interaction:

Perform think-aloud by verbally describing your thought processes while interacting with LLMAs and log your actions. Integrate the ERT to capture emotional responses during interaction. Complete a trust questionnaire after each interaction.

- **Week 1 (4 -8 November 2024).** Carry out the user interaction with the ERT and LLMAs of your choice for **1 academic activity** and **1 non-academic activity**. The minimum interaction duration is 5 minutes each. Design a report template to present raw data (see Section 3.A: Summative Assignment Part 1).
- **Week 2-3 (2-13 December 2024).** Carry out the user interaction with the ERT and LLMAs of your choice for **3 different academic activities** and **3 different non-academic activities,** with each lasting 5 to 15 minutes. Use the **revised** report template to present raw data (see Section 3.B: Summative Assignment Part 2).

3. Goal*:* Evaluate both cognitive (thoughts and actions) and emotional data to estimate their separate as well as combined effect on improving the LLMA's decision-making and user trust.

B. IMPLEMENTATION

1. **Think-Aloud Data Collection**:
   - Use LLMAs to perform activities while verbalising your thought process.
   - Speech-to-Text API: Use any speech-to-text tool (e.g., Google Speech API, Python's `SpeechRecognition` library) to convert verbalised thoughts into text.
2. **Emotion Data Collection:**
   - In the monomodal mode, integrate the speech emotion recognition tool.
   - In the multimodal mode, record your face using a webcam.
   - The ERT analyses the audio/video data and assigns emotion labels in real-time. The emotional data (e.g., timestamps with emotional states) is logged alongside the thought and action data. Alternatively, you can record the interaction and upload to the ERT for analysis after the interaction. Synchronise all the data.

   Emotions to Capture:
   - **Interest/Boredom**: This is critical to understanding engagement during interaction as discussed in *Open-endedness via Models of human Nature of Interestingness (OMNI)* [6].
   - Other emotions: Emotions returned by the ERT selected that you deem relevant.
3. **Action Data Collection**:
   - **Action Data**: Record the actions while interacting with the LLMAs (e.g., questions they type into a chatbot).
   - **Data Logging**: Action data can be logged manually or using simple event logging scripts (e.g., when typing a message into the chatbot, it can be saved in a log file.)
   - **Data synchronization**: Ensure all thought, action, and emotion data are logged with timestamps for later analysis.

C. ANALYSIS

You are expected to perform an analysis based on the three aspects listed below. For each aspect, some ideas and examples are provided. You may choose to address any or all of them, depending on what is most relevant to your data.

1. **Emotional Trajectory Analysis**:
   - **Visualisation**: Analyse how the user's emotional state changes over the course of the interaction. This can be done using graphs or visual timelines that map emotional responses, especially interest vs. boredom at different points during the task. Examples of tools include Python (Matplotlib or Seaborn) or Excel.
   - **Interpretation**:
     - Analyse how the emotional trajectory reflects the user's experience. For instance, a dip in **i**nterest might suggest that the LLMA failed to engage the user effectively at a certain point.
     - Analyse how emotions might align with or diverge from thought processes**.** For example, the user might express frustration verbally (thought), but the ERT could detect rising interest due to the challenge presented by the LLMA.
2. **Thought-Action-Emotion Data Analysis**:
   - **Thought Data Analysis**: Review the thought data to identify patterns in decision-making. Analyse how your thought process guided your actions and what reasoning led to specific choices.

- **With vs. Without Thought Data**: Imagine how the LLMA would have performed if it had access to just the action data (e.g., user's inputs) vs. both action and thought data (e.g., user's reasoning).
- Analyse the following scenarios (or perhaps more)
  - **Thought-Action Mismatch**: Identify where thought and action don't align, highlighting areas where thought data could have prevented a poor decision.
  - **Predicting AI Errors**: Discuss where the LLMA might misinterpret user actions if it doesn't have access to the reasoning behind them (i.e., thought data).

3. **Evaluate the Impact**:
   - Evaluate how thought data might help improve the LLMA's performance.
   - Estimate, without building a model, how thought data could help the LLMA in clarifying ambiguous actions, correcting misunderstandings, or better adapting to user intentions.
   - Evaluate how emotions arising during interaction with the LLMA impact trust in it.

3. ASSIGNMENT REQUIREMENTS

NOTE: As an essential part of the assignment, you are encouraged to verbalise your thoughts naturally while interacting with the applications. However, if you accidentally mention a person's name or share sensitive information, you can redact it by blacking it out before submitting your record templates. Please note that the data will only be reviewed by the module lecturer.

A. SUMMATIVE ASSIGNMENT PART 1 (20%)

1. Tasks

- Complete Week 1 tasks as described in Section 2.A.
- You are required to create a clear and usable **record template** to record the following items:
  - The ERT selected: name, source (URL, if applicable), justification (10%)
  - For the LLMA used for each of two activity types (academic and non-academic):
    - a) LLMA used: name, source (URL, if applicable), justification. (10%)
    - b) The purpose and context of use, be specific about the aim and nature of the activity. (10%)
    - c) Three types of data **synchronised with timestamps**:
      - Thought data transcribed. (10%)
      - Action data logged. (10%)
      - Emotion data returned. (10%)
    - d) Emotional trajectory visualised. (10%)
    - e) Trust measures in both the LLMA and ERT. (10%)

2. Deliverable and Marking Scheme

The main purposes of this assignment are to **design the record template** and to **experiment with the data collection process**. Both the template and the process may be revised for use in Weeks 2 and 3. Note that you are NOT required to analyse the data collected in Week 1.

**** Upload your record template, filled with the data, to Ultra by the deadline ****

The mark distribution is as follows:

- Each item listed above is allocated 10% (=80%) and marked for clarity, thoroughness, relevance.
- Usability of the record template – 10%
- Innovativeness of the record template – 10%

The total score, out of 100%, will be scaled to 20% for the final coursework grade.

B. SUMMATIVE ASSIGNMENT PART 2 (80%)

1. Tasks

- Complete Week 2-3 tasks as described in Section 2.A with the revised data collection process and the record template (cf. Section 3.A.1)
- Perform an analysis for each of the six activities according to Section 2 C: Analysis.
- Compare the analyses across and between activity types.

2. Deliverables and Marking Scheme

- Completed record templates with data collected for three academic and three non-academic activities during Week 2 and Week 3.
- A report on data collection, data analysis, results, discussion, and conclusion (see below)

**** Upload the record templates and report as a zipped file to Ultra by the deadline** **

Format: The report should use the following format:

- Single spacing, font of any style, font size of minimum 11 point, margins of minimum 2cm all round (NB: This layout format is not applied to the cover page)
- Word or PDF file
- Length limit: minimum 2000, maximum: 4000 words (excluding the contents in the record templates, references, and appendices)

Content: Your report should consist of the following sections (the table below). The suggested numbers of pages per section are indicative only. It is at your own discretion to adjust them according to your project. Note that only Week 2 and Week 3 activities are analysed for this report. The total score, out of 100%, will be scaled to 80% for the final coursework grade.

Cover page: Title of your report; Student name; Student number; Date of reporting.

Section 1: Introduction (~ ½ page) (subtotal: 5%)

- Describe clearly the motivation and aim of the project.
- Summarise concisely the LLMAs and ERT selected, the activities where the LLMAs are applied, and key observations derived from the analysis of these activities.

Section 2: Methods (~ 4 pages) (subtotal: 20%)

Describe clearly and comprehensively:

- the technical information of the ERT and each of the LLMAs used. (5%)
- the process of collecting the thought data, action data, and emotion data (10%).
- the ease/difficulty of the process, any challenges and resolutions made (5%).

Section 3: Analysis and Results (~ 5 pages) (subtotal: 55%)

- Availability and relevance of the raw data for analysis (15%)

NOTE: Analyses will NOT be marked unless the templates contain valid and credible data.

3.1 Results of individual activity analysis (30%)

- For each of the three academic and three non-academic activities done in Week 2-3, perform the analysis as described in Section 2C: Data Analysis.
- Each of the analyses will be marked on:
  - Clarity: How clearly the analysis is explained and how easy it is to read and understand.

- ○ Relevance: How well the analysis addresses the issues outlined in Section 2C.
- ○ Rigour: How convincingly the analysis is supported by the raw data and grounded in sound methods.

3.2 Results of comparisons across activity types (10%)

Compare (i) across the three academic activities; (ii) across the three non-academic activities; (iii) between the two sets of academic vs. non-academic activities to identify any specific patterns indicating **the effect of activity type** on the relationships among thought data, action data, and emotion data. Comparisons will be marked on their clarity, relevance and rigour (cf. Section 3.1)

Section 4: General Discussion (~ 1.5 page) (subtotal: 15%)

In your discussion, reflect on the following questions based on your practical experiences in this project. Argue whether your insights align with or contradict the relevant Human-AI Interaction (HAII) frameworks covered in the lectures:

- How useful and reliable is think-aloud data for AI in modelling human reasoning? Comment on the potential of the Thought Cloning approach.
- How useful and reliable is emotion data, especially interest/boredom, for AI in modelling human notion of interestingness – the OMNI approach?
- How stable is trust in AI applications? Does familiarity with an AI system enhance or undermine trust in it?
- What are the ethical implications if AI is able to fully mimic human reasoning?

Your responses to these questions will be evaluated based on their clarity and how convincingly they reference relevant HAII frameworks.

Section 5: Conclusion (~ ½ page) (subtotal: 5%)

Present a succinct summary of the project's work as a whole and an outlook for the future work.

References

Appendices

REFERENCES:

[1] Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215.

[2] Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, 5(3), 178-186.

[3] Fan, Y., Rakovic, M., van Der Graaf, J., Lim, L., Singh, S., Moore, J., ... & Gašević, D. (2023). Towards a fuller picture: Triangulation and integration of the measurement of self-regulated learning based on trace and think aloud data. *Journal of Computer Assisted Learning, 39*(4), 1303-1324.

[4] Hu, S., & Clune, J. (2024). Thought cloning: Learning to think while acting by imitating human thinking. *Advances in Neural Information Processing Systems* (NeurIPS), 36.

[5] Ostrovsky, T., & Newell, B. R. (2024). Verbal reports as data revisited: Using natural language models to validate cognitive models. *Decision*.

[6] Zhang, J., Lehman, J., Stanley, K., & Clune, J. (2023). OMNI: Open-endedness via models of human notions of interestingness. arXiv preprint arXiv:2306.01711.