Human AI Interaction- Human AI Trust

Sabrina Afrine Sathi
Master of Cyber Security
The University of Adelaide
Adelaide, South Australia
sabrinaafrine.sathi@student.adelaide.edu.au
Supervisor: Olaf Maennel

Abstract – *The increased focus on trust among users in AI-enabled systems indicates its importance as a key driver of adoption, mandating an upgrade diverting attention from solely technological solutions and toward human-centric methods connected with the principles of human computer interaction (HCI) [3]. Artificial Intelligence (AI) is integrated in human decision-making processes which emphasizes the need of understanding trust dynamics; nevertheless, empirical research comes across challenges due to inconsistent experimental methods of evaluation [1]. The study explores the trust dynamics in AI supported systems by developing and executing a generalized evaluation framework with the utilization of Large Language Models operating as judges (LLM-as-a-judge). With the growing implementation of AI in areas of important decision-making, the concept of human trust and its measure should gain the utmost priority; however, existing evaluation procedures are problematic due to inconsistency in experimental designs and the absence of a unified definition of trust. This paper addresses these difficulties through the multi-phase approach that incorporates human-centered survey research as well as automated evaluation methods. A survey with broad demographics and profiles investigated the behavioral patterns associated with the use of AI, factors of trust, and concerns in real-life applications. Three main trust-affecting factors have been recognized in the research: Socio-ethical issues (fairness, transparency), technical specifications (accuracy, explainability) and user demographics (experience, cognitive styles). And the fundamental contribution is a new framework of LLM as a judge evaluation. It provides evidence-based metrics to evaluate human-AI trust and allows reliable systems and improve cooperation in higher risk use cases.*

Key Terms – Artificial Intelligence (AI), technology, human, decision-making.

Github Link- https://github.com/Sabrinaafrine/Llm-as-a-judge-evaluation-framwork

INTRODUCTION

Artificial intelligence (AI) is becoming more prevalent in everyday life, augmenting technology such as smart devices, fitness trackers, self-driving cars, and social networking platforms [7]. AI-assisted decision-making has become critical in high-risk sectors such as healthcare, recruiting, and criminal justice, where results are unpredictable, and repercussions are severe [1]. Many AI systems acquire autonomy and human characteristics, making them social agents [7]. Trust in Artificial intelligence (AI) systems is critical for successful human-AI cooperation, but it presents substantial theoretical and practical challenges [6]. Moreover, their widespread adoption raises ethical issues, notably about biases, the loss of human agency, and the anonymity of data-driven algorithms. Trust in AI systems holds significance for increasing satisfaction with systems and promoting explainable AI (XAI), that is intended to balance trust with system integrity and consistency [7]. Theoretical issues emerge from contradictory definitions and overlapping terms associated with related words such as dependency, trust and calibrated trustworthiness, resulting in unclear operationalizations [6][8][9][10]. For example, trust indicates a psychological attitude, while reliance evaluates changes in behavior after AI support [10][11]. Furthermore, the underemphasis on rational mistrust (doubt toward distrustful AI) hinders integrative knowledge, since XAI strives not only to increase trust but also to explain distrust in ineffective systems [12][13]. According to surveys, confidence in AI is still low among U.S. consumers and organizations, with just 16% trusting AI in medical diagnoses, 13% in self-driving vehicles, and 4% in HR duties [14].

This research analyzes the evaluation of human trust in artificial intelligence (AI) systems, highlighting its significance for AI adoption and engagement. Trust is described as an attitude that expresses faith in an agent's capacity to assist accomplish objectives in the face of uncertainty and vulnerability. Because trust is a psychological concept, it is tested using questionnaires that enable studies to compute a subjective trust score for statistical analysis [15]. Furthermore, to evaluate AI models to find out the limitations in matching human preferences, this study demonstrates using powerful large language models (LLM) like “Qwen/Qwen2.5-7B-Instruct”, “google/gemma-2-9b-it:nebius” and “meta-llama/Llama-3.1-8B-Instruct:cerebras” as evaluators of other models on open-ended tasks, also known as LLM-as-a-judge. It examines the advantages and the drawbacks of implementing LLMs as judges, finding problems with positions bias, verbosity bias, self-enhancement bias, as well as limited reasoning power. It also determines the quality of the concordance of the LM-based evaluation with human judgment proposing two criteria, such as MT-bench, which has multi-turn formulations, and Chatbot Arena, a crowdsourced evaluations platform. Human evaluation is the most reliable way of measurement of human preferences, but it is usually time consuming and costly. In a way to ease up, we explore the idea of replacing people judges with Machine learning through utilization of cutting-edge large language models such as GPT-4 and Gemini 2.5 pro. [49]. In this report, a different evaluation framework is used by including the concept of human preference benchmarks and introduces an automated evaluation technique that can be scaled to use LLMs as judges, or LLM-as-a- judge, and demonstrates the improvement in the trust in AI [49].

AIMS AND CHALLENGES

The tables below outline the aims and challenges of this project:

Aims	Challenges
Develop an effective framework capable of proficiently addressing human inquiries and deliveries precisely.	Choosing the most appropriate model that would meet a balance between performance, expense, precision and flexibility to various types of questions.
Enhance complete understanding and efficiency of AI systems to increase user confidence and promote improved decision-making.	Data availability and quality, including the difficulty of getting enough, high-quality datasets that correctly represent the target population or issue space.
Improve the AI solution's overall efficiency, accuracy, and usability, making it more trustworthy and useful for humans.	Ethical or security problems include safeguarding user privacy, eliminating bias in AI decision-making processes, and dealing with possible technological exploitation.
Run a survey on the questionnaire to gather data on human AI interaction for the dataset.	Insufficient survey replies slow down the collection of accurate statistics.
Comparing AI-generated answers with open-ended questions with the help of human judgment to understand how AI-generated content can fit in human expectations.	Open-ended evaluation necessitated an even greater reach and contact in the process of obtaining participants than was envisaged. The narrow sample size rendered it hard to exhaustively conclude or make conclusions that are general.

LITERATURE REVIEW

1. Defining AI Trust

Trust in AI is a broad idea centered in the expectation that an AI system would operate in the user's best interest, especially in situations of uncertainty, crisis, or inadequacy [6][12][17]. It goes beyond technological dependability and includes psychological, ethical, and social components. Philosophically, trust is linked to moral connections; in the field of social science, it expresses a mindset of non-hostile conduct [6]. Economic frameworks usually construct trust using game theory, with the concept stressing strategic reciprocity [6]. Most formulations revolve around the notion that trust originates in circumstances when people must depend on AI despite insufficient awareness of its decision-making processes [6][17]. Transparency, comprehensibility and accountability are fundamental elements that help people perceive the logic of AI and avoid biases [12][17][19]. Trust depends on technological performance (accuracy, security), fairness, and compliance to principles of ethics such as privacy and non-discrimination [12][19][20]. More importantly, mistrust is not only the lack of trust, but also active doubt about an AI's conduct, frequently owing to perceived hazards or past experiences [4][6]. Trustworthy AI systems need strong validation, human monitoring, and proper alignment with user demands throughout their lifespan [12][19]. This interaction of dependability, ethical approach and user experience describes trust to be both a psychological state and a quantitative attribute of artificial intelligence systems [6][19][20].

2. Building and measuring confidence in AI

Trust in AI systems originates from the interaction between a system's built-in trustworthiness and its users' impressions of the system's abilities. The fundamental paradigms recognize three key elements of trustworthiness: ability (technical competence), kindness (alignment with user objectives), and integrity (adherence to ethical norms) [10]. These characteristics are evaluated using visible indicators like as performance measurements, transparency methods, and design elements that indicate dependability or ethical alignment. However, there is a significant contrast between true trustworthiness (objective system features) and perceived trustworthiness (subjective user ratings). For example, a user may trust an AI based on reliable job performance, irrespective of whether internal prejudices or security concerns exist. Recent study focuses on warranted trust, which is supported by provable trustworthy traits rather than superficial considerations like anthropomorphic design. Achieving this balance requires constant review across the AI lifecycle—from confirming data used for training integrity to track post-deployment user interactions [19][21]. While technical criteria (accuracy, robustness) establish the basis, human-centered elements (explainability, fairness) eventually decide whether trust becomes a stimulus for implementation or an obstacle to AI's social integration [8][22].

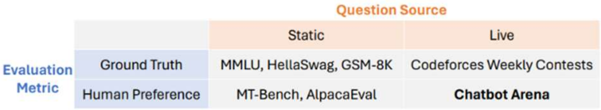
3. Distrust in AI

The difference between warranted and unwarranted trust in AI is based on matching human perceptions to a system's real trustworthiness. When an AI system is clearly untrustworthy (for example, owing to biased outputs or low accuracy), trust is inappropriate, and active skepticism is justified to avoid detrimental dependence. This brings into query the conventional perception of explainable AI (XAI) as just a tool for boosting trust; alternatively, XAI should attempt to calibrate user expectations, raising confidence in competent systems while instilling mistrust in faulty ones [6][23][24]. Current research is too focused on trust, ignoring distrust's role in avoiding heavy reliance on untrustworthy AI. This contradiction continues despite evidence that mistrust promotes important behaviors—such as rejecting biased algorithms or demanding accountability—which trust-centric theories cannot explain [25]. Furthermore, human-AI interaction paradigms must utilize two-dimensional models to assess and treat these components individually, guaranteeing that users do not blindly accept defective systems or reject helpful ones owing to uncalibrated skepticism.

4. Guidelines for Human-AI

Modern Human-AI interaction standards prioritize transparency, comprehensibility and human-centered design as fundamental pillars for promoting user trust, with at least 30% of ethical frameworks expressly valuing trust as a basic concept [10][26]. Although these guidelines are based on practitioner observations and empirical investigations, their actual implementation sometimes lacks clarity on how particular design choices—such as interpretable decision outcomes or adaptive interfaces—directly affect trust dynamics [27][28][29]. The newly formed frameworks, such as Accenture's, emphasize human-centered design as vital for creating trust, pushing for iterative user interaction to address ethical problems including bias mitigation along with accountability [30]. Despite this, the field encounters challenges in demonstrating which guidelines consistently boost trust calibration, particularly in choosing between deceptive compliance (e.g., presenting excessive technical information). and effective trust-building. Future guidelines must use evidence-based methodologies from cognitive science and human aspects research to enhance recommendations, making sure they satisfy both warranted trust (supported by system dependability) and warranted mistrust (due skepticism towards defective systems) [29][31]. This process will need a thorough examination of how user variety, context-specific dangers, and longitudinal interactions influence trust growth beyond universally applicable prescriptions.

5. LLM Benchmarks



		Question Source	
		Static	Live
Evaluation Metric	Ground Truth	MMLU, HellaSwag, GSM-8K	Codeforces Weekly Contests
	Human Preference	MT-Bench, AlpacaEval	**Chatbot Arena**

Figure 1. LLM Benchmarks [56]

The benchmarks of LLM can be broadly classified into the static and dynamic ones. Static, ground-truth-based benchmarks, based on multiple-choice or pre-determined answer formats, are the most common as they verify such skills as language understanding, math, coding, and reasoning. Main ones are MMLU [50], HellaSwag [51], GSM-8K [52], BigBench [53], AGIEval [54], and HumanEval [55], shown in figure 1. Benchmarks are also available aimed at safety (e.g. ToxicChat) and evaluation suites in general, such as HELM [56].

In addition to these closed-ended tests, other benchmarks may use open-ended questions to which human judgments need to be obtained, typically retrieved by experts or crowdworkers. More recently, there has been a trend of slightly imitating the human judgment criteria as assessed using LLMs such as GPT-4 (as with MT-Bench and AlpacaEval). Besides static assessment benchmarks, live benchmarks with current questions of exams, coding contests or user interaction are becoming more popular [56].

These live interactions have in some studies been used to train models via reinforcement learning based on human feedback, however more often such work is internal within a particular organization. In response to the absence of open human-in-the-loop evaluation this paper uses Chatbot Arena, the first large-scale, geographically diversely distributed, openly accessible human-bots benchmarking platform based on actual human interaction [56].

6. Limitations in Static Benchmarks

There are several shortcomings of static benchmarks, including data contamination, performance saturation, overfitting, and lack of alignment with human preferences [58]. Research projects such as DynaBench [57] have brought these concerns into the spotlight and propose live, human-in-the-loop evaluation approaches to enhance the classical NLP benchmarking. Potentially developing this concept further, the system proposed here draws upon the concept of including real-time human interaction but is more specific in its nature, even with the added aspect of introducing the technology at a greater user level [56].

7. Model Ranking

Different aspects of ranking systems have been studied in statistics, including online experiment design, probability models and rank elicitation. Elo rating system is also used to rate LLMs [59] [60].

8. Human preference Dataset

In understanding the value of human preferences, multiple datasets have been developed to measure them, including OpenAssistant, HH-RLHF, LMSYS-Chat-1M, and synthetic datasets, including UltraFeedback and Nectar. Despite LMSYS-Chat-1M, a dataset previously published by the authors, being a result of crowd-sourcing, it contains solely conversation data without annotations of the human preferences, which eliminates its usability in ranking problems. Conversely, the focus of the current paper is on the analysis of human preference data about ranking assessments [18][56].

Survey Ethics

This project is based on a survey that required participants. This has created a concern for ethical considerations. In general, all organizations have their own set of rules for conducting surveys or experiments on human subjects. However, the basic ethical guidelines are broad, fundamental judgments that govern and justify ethical standards and evaluations of human conduct [32]. Three fundamental sociocultural ethical concepts appear in the setting of human subject research:

1. Respect for Persons: This concept highlights recognizing persons as independent individuals while safeguarding those who possess restricted freedom of choice [32].

2. Beneficence: entails the commitment to do good, maximize potential benefits, and limit potential damage to study participants [32].

3. Justice: This concept emphasizes justice in allocating research rewards and responsibilities, to make sure no group is unduly oppressed or left out of possible advantages [32].

The purpose of these principles is to serve as the ethical underpinning for doing research with human subjects, directing how researchers should treat and safeguard those participating. Likewise, the University of Adelaide has an ethical form that outlines the ethical terms to protect the student and other individual from any harm. Every survey must be conducted after the NHMRC's form is approved by the university's ethics committee. As the survey conducted is human research and it complies with the NHMRC guidelines and falls under "Low risk research." The research ticks all the boxes that are included in the category by Lower Risk Human Research Ethics Committee (LRHREC) [16]. Furthermore, to respect respondents' privacy, participation in the survey was entirely voluntary and the email address was collected for authenticity, however the email address was removed from the dataset during the cleaning process.

Progress and Plan

The progress of the project is as stated in the timeline below. The plan in week 7-8 ticks all the boxes. The hugging face discord channel is joined. The survey questions are prepared and shared among the peers, forums and discord channels.

Timeline

Trimester 1	Week 1-2	Project idea brainstorm
	Week 3-4	• Outline research objectives and hypotheses • Understand the project and find related research papers • Develop a work breakdown structure (WBS)
	Week 5-6	• Write a research proposal • Sumit the research proposal
	Week 7-8	• Prepare questionnaire • Join Hugging Face discord channel • Conduct survey for data collection • Write midterm report

	Week 9-10	• Identify existing datasets as well as research gaps • Create, organize and clean data • Generate synthetic data
	Week 11-12	• Setup pipeline and frameworks • Prompt LLM to judge and evaluate case studies. • Test the data pipeline to confirm it is working properly. • Write a final report and include the process, dataset, pipeline setup, methodology and lessons learned. • Submit the report
Trimester 2	Week1-2	• Review and finalize the proposed framework and pipeline • Set up data transformation and integration tools. • Current framework analysis
	Week 3-4	• Examine the curated dataset to find AI trust trends and patterns. • Use statistical tools to determine bias and trustworthiness.
	Week 5-6	• Run LLM analysis to gain insights into AI trust.
	Week 7-9	• Compile the findings of the analysis and LLM answers. • Determine key findings on AI trust and any data biases.
	Week 10-12	• Write a report outlining the process, analysis and results • Discuss recommendations for enhancing trust in AI based on the results reported. • Review the report for readability and its accuracy and submit.

METHODOLOGY

This section outlines the approach of understanding what kind of biases prevent trust in AI and how this can be investigated [49].

Within the original configuration, the model answers were created to all the 81 MT-bench questions with the help of six varied language models. The two groups of people carried out judging:

- LLM Judges: Auto grading of big language models.

- Expert Human Labelers: 58 well-qualified human annotators (most of them are grad students), each with at least 20 multi-turn questions evaluated randomly. This resulted in a set of around 3,000 human-labeled votes [49].

The study also sampled 3,000 single-turn evaluations of the large-scale Chatbot Arena (30,000+ interactions), which also had wide range of language models. There was judging of both LLMs and a group of over 2,100 distinct human participants who were crowdsourced [49].

Furthermore, as the AI is advancing rapidly, the dataset with 81 MT-bench questions were used in older versions of AI models in the Chatbot Arena. Therefore, this study extended the methodology developed to evaluate the language models with the already established corpus of MT-bench questions.

The methodology of the evaluation is a four-section pipeline that has a structured assessment strategy to guarantee a thorough and dependable outcome of the evaluation. In phase one, the data collection and preprocessing operations involve such actions as severe validation of question-answer pairs, various criteria selection and customization, initialization of the judge model with authentication of health status, and rate limiting parameter optimization to demonstrate the successful API communication. Phase two deploys the operational multiple-judge evaluation by guarding parallel evaluation implementation methods in view of rate limitation restriction and dependability necessities. This step produces structured evaluation queries containing sufficient details of criteria descriptions and context and then analysis of the response and validating the respective answers to guarantee standard response format. In phase three various aggregation and analysis of evaluation outputs takes place, using complex calculations of inter-judge agreement, by ensemble scoring and its corresponding confidence intervals, extensive bias analysis and correction processes, and automated production of recommendations because of statistical analysis outcomes. The fourth phase involves full reporting and visualization such as statistical significance testing, performance trend analysis on many dimensions of evaluation and creation of interactive visualization dashboards to explore results in detail.

The framework uses strict statistical analysis techniques to obtain credible evaluation outcomes that can be interpreted. The Bootstrap confidence interval estimation involves resampling of 1000 samples which are applied in producing confidence limits of agreement scores estimated at 95 percent and offer sound reliability measures of effectiveness taking into consideration the variability of sample assessments. Such method assists practitioners in determining the statistical significance and their reliability limits of evaluated results. Both Spearman rho rank correlation coefficients of ordinal evaluation data and Kendall tau correlation of tie-robust correlation assessment are used in correlation analysis and can give a complete evaluation of cross-judge and ranking agreement.

The MT-Bench integration component incorporates a complete data processing pipeline which starts by loading the datasets of human judgment of the LMSYS MT Bench repository, the integration of the GPT-4 data of pairwise comparison and their enrichment with question metadata such as automatic labeling on categories. The system goes through the processes of vote normalization that provide different types of tie formats and consistency in data representation across all sources of evaluation. The pairwise comparison approach adheres to the consistent ordering of models via the alphabetical sorting, uses extensive normalization of vote labels to transform different forms of ties into uniform standard labels to be referred to as ties, and supports multiple-turn type evaluation scenarios, including both first-response evaluations (Turn 1) and subsequent assessment after the response is given (Turn 2). The human expert judgments as the reference of ground truth are applied in reference-based evaluation and compute the percentage rate of agreements of the human established judgments with an applicable interval of confidence and thoroughly provide the ranking correlation appraisal through Spearman and Kendall correlation coefficients to determine accuracy of the automated evaluation methods.

Additionally, the evaluation framework incorporated multi-model to help review the quality of Large Language Model (LLM) answers on several dimensions. The evaluation process was conducted via a formal prompt engineering mechanism, and every evaluator model had the same evaluation prompt that included the original question, the answer to be judged, and the in-depth descriptions of the evaluation criteria. All the criteria were rated on a 0-10 scale using standardized descriptions so that they would be consistent across raters. Inter-evaluator agreement was evaluated using methods comprising agreement scores (normalized standard deviation), agreement ratios (proportion of matching evaluator pairs below threshold) and tie ratios (frequency of nearly identical scores) and Krippendorff's alpha for reliability calculation. This approach produced detailed recommendations depending upon overall performance with absolute scores as well as inter-evaluator agreement patterns and the recommendations reflected the subtle interpretations of the approach that took note of both the quality of responses and the reliability of the assessment. The multiple faceted models can be used to perform robust evaluation that can be applicable in individual level response assessment and in the large-scale batch evaluation.

BENCHMARK DESIGN

With the recent developments in large language models (LLMs), AI assistants have been shown to exhibit the symptoms of artificial general intelligence, as they could fulfill a variety of tasks, encompassing simple conversation and writing, too complicated coding. Assessing these broad skills is, however, getting very complex. Although a good number of LLM benchmarks have been created, they usually involve only a limited set of evaluations, i.e. tasks with short and closed-ended frame responses that do not comprehensively evaluate a wide range of task capabilities such as open-ended, multi-turn, and following instructions of the new chat assistants [49]. The existing benchmarks can be divided, in general, into three categories:

- Core-Knowledge Benchmarks- These tests involve zero- or few-shot conditions in assessing early skills, and there are well-defined solutions that can be judged automatically like MMLU [50], ARC, GSM-8K [49] [52].
- Slightly-Open Ended Tasks Instruction-Following Benchmarks- These are provided to present semi-open participated activities, which are appropriate to model that learn how to do what they are told to do for example – Flan [49].
- Conversational Benchmarks- Such benchmark problems of the multi-turn dialogue mirror actual chatbots scenarios, although the complexity and diversity of their questions usually do not keep the current best models on their toes, for example- CoQA [49]

Most importantly, a key feature that tends to be absent in such benchmarks is considering human preferences directly measured by the usefulness and subtlety of chatbots in open-ended conversations. This thesis therefore provides two new benchmarks that are aimed on one hand to measure usefulness as seen by human users and, on the other, to benchmark the main advantages of state-of-the-art LLMs [49].

MT-Bench is a benchmark created to test multi-turn conversation and instruction-following skills of language models. It includes 80 well-designed and high-quality multi-turn prompts, specially curated to cover real-world situations in the limelight and test the feat of advanced LLMs. The objective is to point out variations among models by using complex interactions rich in contexts. To have a thorough assessment, the benchmark deals with eight familiar categories of prompts- writing, roleplay, extraction, reasoning, math, coding, knowledge I (STEM), knowledge II (The humanities/Social science). There are 10 handcrafted multi-turn questions in each category and as such, the test set is well rounded and diverse. These questions aim to resemble easy and challenging situations that will determine how multidimensional a model is, whether it has any internal consistency and is ready to answer in a sustained conversation. [49].

Chatbot Arena is a crowdsourced benchmarking system that tests language models in an anonymous model-versus-model contest of last man or model standing. The interaction in this platform involves a user asking a question and then getting replies by two unidentified chatbots. They then cast their votes on the answer they like more, and the name of models will be displayed after voting. It is not based on pre-determined questions, as in the case of traditional benchmarks; rather, it applies to a wide range of queries and user preference across a great variety of real-world questions, which show authentic and free-model usage scenarios [49].

This report presents a sophisticated multi-model assessment infrastructure that assesses Large Language Models (LLMs) through two modular components, which include real-time evaluation based on HuggingFace and benchmark-based evaluation capabilities.

1. Architecture of Multi-Modal Evaluation

The real-time evaluation system allows a live assessment on question-answer pairs provided via live model APIs and has interactive evaluation sessions with multi-criteria assessment across seventeen different dimensions of evaluation. It allows ensemble-based scoring and provides details of the agreements and is featured with clear CLI interface where feedback and revision can be done in real time.

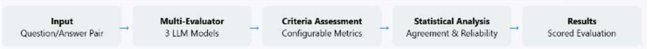


Figure 2: System Architecture Flow

In addition to the real time system is a detailed benchmark analysis platform which conducts a retrospective analysis of the known judgments of the MT-Bench. This system is based on pairwise comparison technique and implemented on systematic model ranking with specialized evaluation criteria adapted to systematic model ranking, particularly to science domain assessment. The benchmark analysis module offers extensive bias identification capability and conducts regulated consistency review of the wide range of evaluation dimensions and substantiates the overall evaluation of results and the performance patterns of models.

2. Criteria of Evaluation Framework

The evaluation framework has seventeen separate assessment criteria that are structured following rigorous categorization into core and specialized rubric categories, which result in a broad coverage of the dimensions of response quality. The benchmark criteria represent the essential requirements of the quality of responses such as the accuracy that quantifies the factual correctness and truthfulness of the information expressed. Helpfulness measures the usefulness of responses in responding to certain user needs and requirements, whereas clarity examines communication efficiency and overall understandability of the content that is generated.

The framework goes beyond mere basic criteria to deeper dimensions of assessment, which utilizes more accurate faculties of the quality of responses. Completeness is used to determine the extent to which the answers are covered adequately against the initial scope of question, whereas coherence is assessed based on logical consistency and orderly presentation. Creativity and innovation are also evaluated where it is relevant to the situation, but this is balanced against demands of conciseness which imply that the coverage must also be thorough. Other focused requirements are empathy, as an alternative to measuring emotional intelligence and relevance, actionability as an alternative way to measure practicality, and depth as an alternative strain to measure analytical competence.

3. **Specific Category (Science)**

The framework also includes high-end evaluation abilities that would specifically be developed to evaluate scientific content using state-of-Art-feature extraction and analysis methods. The system measures the density of technical terms by automatically detecting and quantifying scientific words as compared to the total length of contents and gives a clue about the technicality and suitability of scientific explanation. The citation pattern recognition algorithms search through academic formatting rules and mention structures, and the quantitative information analysis identifies the availability and correct use of quantitative data, measurements and statistical data.

Science specialization goes further by including methodological content assessment whereby the system assesses the use and quality of the discussion of scientific methodology, description of experiments and their experimental procedure and method of analysis. The framework also involves uncertainty awareness assessment whereby high-quality communication of science is meant to take note of limitation, confidence intervals, and areas of continuing research/debate. These special abilities allow the framework to offer high dimensionality rating of scientific content quality that transcends the general-purpose evaluation indicators.

LLM-AS-A-JUDGE EVALUATION SYSTEM

Initial experiments based on MT-bench and Chatbot Arena are based on human ratings which are also costly and time-consuming to acquire [18]. To resolve such, this report suggests establishing a more scalable, automated method of evaluation. Because most questions in MT-bench and Chatbot Arena are open-ended, with no easy reference answers, programs using regular metrics that compare with references like ROUGE or BLEU (that measure similarity with reference answers) cannot be helpful.

As LLMs continue to develop rather quickly, curiosity has arisen concerning the question of whether they can be used as reliable evaluators themselves, effectively putting human evaluators out of work in most evaluation operations. The next direction is turning to the very idea of using LLMs as judges that is, simply asking LLMs to score the output of a chatbot and comparing them with the preferences of humans.

Large language models (LLMs) as judges can be used in three ways to grade the replies generated with the help of AI. All these techniques may be applied individually or in combination:

1. Pairwise Comparison: LLM judge is presented with a question and two answers which differ from each other, determining which of them is better or in case of a tie. Although it works, the method is less scalable and more models are to be compared, as the number of pairings grows exponentially [49].
2. Reference-free Grading: In this case, the LLM judge responds in terms of the points per response. Such a method can be scaled more easily but it could not be effective in showing subtle differences between similarly performing models and absolute scores might change in case the judge model is revised [49].
3. Reference-Guided Grading: In the case of a reference answer (as in math problems), the response is given compared to the reference one and graded as such by the judge of LLM [49].

Furthermore, the framework in this study uses LLM-as-a-judge for the process of evaluation that is an advanced type of ensemble method where the evaluation is done with a combination of several large Language Models as separate judges to reduce any single model biases and lead to reliable evaluation overall. The main set of evaluators includes three well-chosen models: Qwen/Qwen2.5-7B-Instruct for multilingual reasoning and mathematical assessment, Google/Gemma-2-9B-IT for instruction-tuned evaluation and Meta-Llama/Llama-3.1-8B-Instruct for reasoning and open-source evaluation. The system also has an adapted version of an enhanced Krippendorff alpha used in ordinal evaluation data. The edge cases that are handled correctly are the condition of perfect agreement as well as the situation of binary outcome reliability assessment by using a variance-based method of solid evaluation of reliability through multi-evaluators. The framework divides the strength of consensus into five groups (very strong consensus requiring scores on agreement above 0.9 and rates on agreement above 0.8), strong, moderate, weak consensus and very weak consensus with each one of the strengths being calibrated at different levels depending on empirical assessment data. The framework also includes multifactor bias-detection systems detecting and measuring different types of systematic evaluations bias. Position bias analysis the tendency to favor model A or B responses is analyzed, and order effects (and other ordered patterns of preference) may suggest weaknesses on the part of an evaluator. The system derives bias tendency coefficients of every single evaluator model, and these factors are taken into consideration by the reliability-weighted scoring algorithms.

Individual bias tendency coefficients, consistency score weightings, and tie preference modeling are used carefully to model and consider model-specific bias characteristics. Some of the mitigation methods of bias ensemble voting that automatically corrects any bias, weighted score ensuring that the binary scoring system is associated with the reliability based on different types of content, and cross-validation that checks the consistency of the method by using repeated independent evaluation sessions. Such methods make sure that systematic biases have been detected and measured and have been corrected in overall assessment scores.

TOOLS AND TECHNOLOGY

1. Technical Stack

The framework is based on a solid technological infrastructure that uses Python 3.8 as the main programming language in the development, because it has a large community of scientific computing and machine learning libraries. OpenAI SDK comes with smooth API integration with the HuggingFace Router service that makes it possible to work with various state-of-the-art language models via a single API. NumPy and Pandas create the analog of solving of numerical processing of data or a base of manipulations with data, which allows working with large amounts of data on evaluation, as well as with sophisticated statistical computations in an effective way. The capabilities of the statistical analysis are given by SciPy that allows conducting advanced statistical operations such as bootstrap, correlation analysis, and hypothesis testing. The HuggingFace Datasets library offers a way to quickly load and manipulate the MT-Bench evaluation data and generally offers a set of means to access and preprocess larger quantities of data used during the evaluation.

2. Rate Limiting & API Integration

API integration is implemented by Real-Life client configuration that contacts the HuggingFace Router service with the help of OpenAI compatible endpoints. The system also has sophisticated rate limiting features such as per-model request throttling with customizable minimum delay (setting defaults to 1-second inter-request delay), exponential backoff rebalance procedures to automatically recover rate limits situations, and parallel request handling that used threading requests to lock and unlock so as to avoid the risk of race conditions and achieve optimum API usage.

3. Data Processing Design

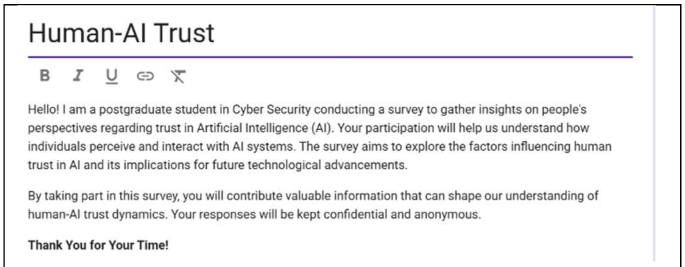
The data processing structure is designed to employ an intensive nested dictionary structure by evaluation turn, and every turn consists of mappings of question-model pair tuples to judge specific sets of votes. This structure allows effective access to evaluation information and ensures clear distinction between various evaluation tasks and judge viewpoints. Processes of model normalization guarantee consistency in the alphabetical ordering of model similarities, consistent label formats for casting votes and canonical forms of judge names and automatic category assignment based on question content and metadata. The system also performs extensive data validation processes that check format consistency of input, find and process missing or malformed evaluation data and present detailed error reporting related to data quality problems. Complex indexing and caching procedures optimize data accessing patterns in great scale evaluation datasets and minimize resource demands during processing of massive evaluation corpora using memory efficient data structures.

4. Interactive System Capabilities

The interactive system includes an advanced command-line interface that in addition to interactive evaluation sessions also allows programmatic execution of large numbers of evaluations in a batch mode. Real-time scoring features have such properties as live study progress monitoring, fixed history of sessions to store evaluation performance of a user, and ability to create fully customizable criteria selection so that a user could pay attention to only the aspects of the evaluation relevant to his or her use case. A batch processing ability is also provided, which allows and evaluating several question-answer pairs simultaneously, parallel and sequential processing strategies.

SURVEY QUESTIONNAIRE

A survey was designed to collect data on AI use trends and the perceived influence on numerous parts of everyday life. Demographic characteristics (age, country, employment) were asked, as well as AI-specific questions like the AI services or tools used, frequency of AI usage, rating the usage impact, etc. The survey was published online to a varied sample of respondents, using standard methods for factual and quantitative research [32].



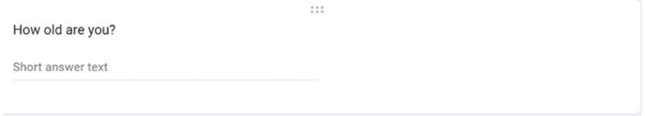
Human-AI Trust

Hello! I am a postgraduate student in Cyber Security conducting a survey to gather insights on people's perspectives regarding trust in Artificial Intelligence (AI). Your participation will help us understand how individuals perceive and interact with AI systems. The survey aims to explore the factors influencing human trust in AI and its implications for future technological advancements.

By taking part in this survey, you will contribute valuable information that can shape our understanding of human-AI trust dynamics. Your responses will be kept confidential and anonymous.

Thank You for Your Time!

Figure 3


How old are you?

Short answer text

Figure 4

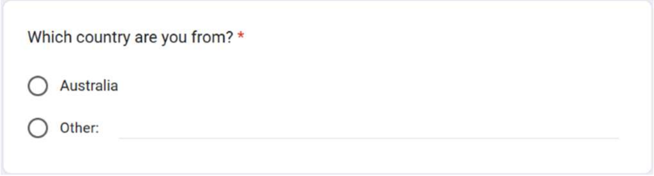
What is your profession *

□ Student

□ Employed

□ Unemployed

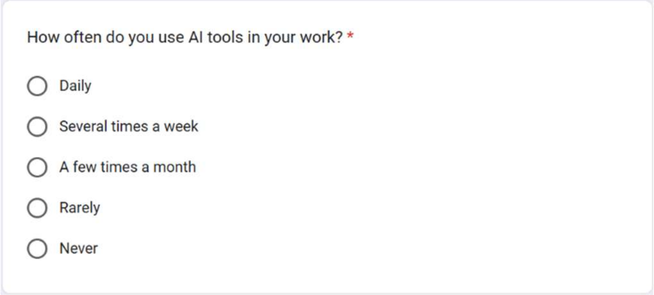
Figure 5


Which country are you from? *

○ Australia

○ Other:

Figure 6


How often do you use AI tools in your work? *

○ Daily

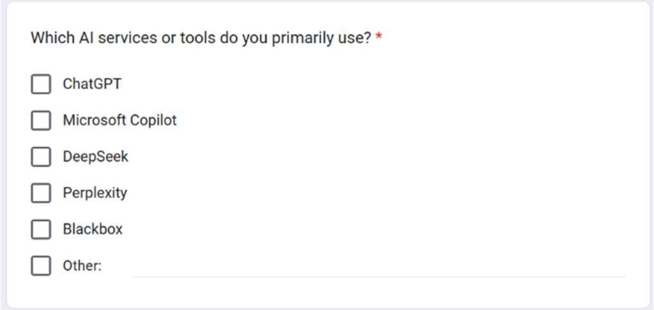
○ Several times a week

○ A few times a month

○ Rarely

○ Never

Figure 7


Which AI services or tools do you primarily use? *

□ ChatGPT

□ Microsoft Copilot

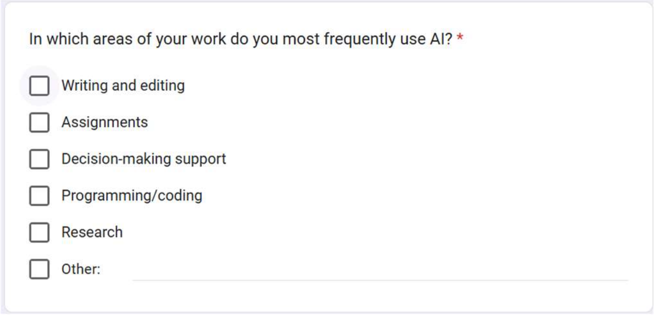
□ DeepSeek

□ Perplexity

□ Blackbox

□ Other:

Figure 8


In which areas of your work do you most frequently use AI? *

□ Writing and editing

□ Assignments

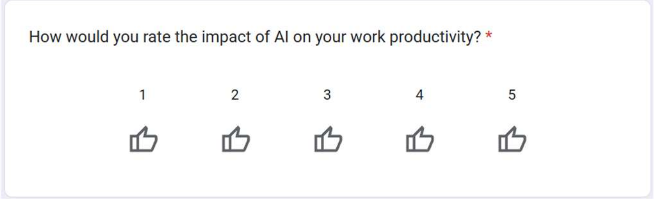
□ Decision-making support

□ Programming/coding

□ Research

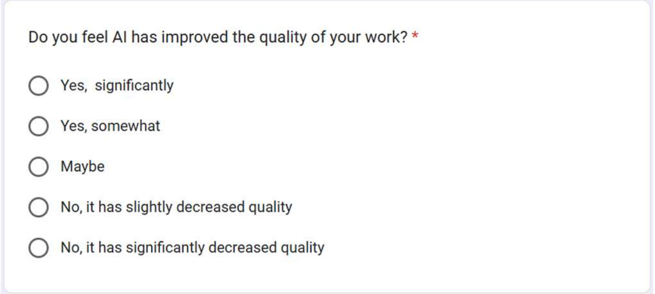
□ Other:

Figure 9


How would you rate the impact of AI on your work productivity? *

1	2	3	4	5
👍	👍	👍	👍	👍

Figure 10


Do you feel AI has improved the quality of your work? *

○ Yes, significantly

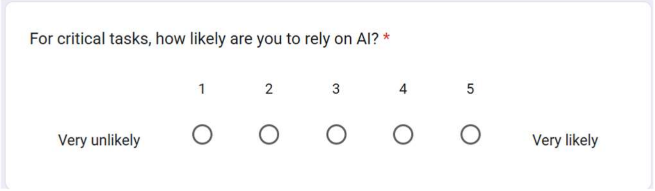
○ Yes, somewhat

○ Maybe

○ No, it has slightly decreased quality

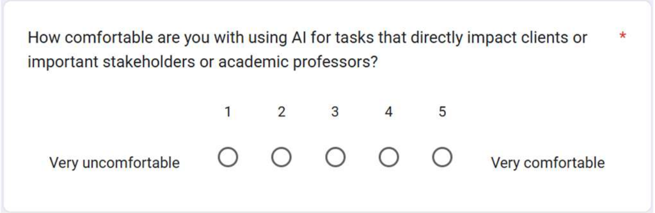
○ No, it has significantly decreased quality

Figure 11


For critical tasks, how likely are you to rely on AI? *

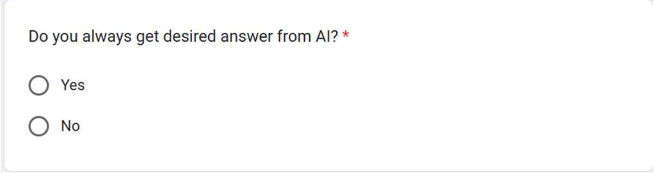
	1	2	3	4	5	
Very unlikely	○	○	○	○	○	Very likely

Figure 12


How comfortable are you with using AI for tasks that directly impact clients or important stakeholders or academic professors? *

	1	2	3	4	5	
Very uncomfortable	○	○	○	○	○	Very comfortable

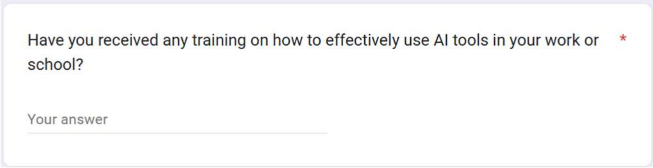
Figure 13


Do you always get desired answer from AI? *

○ Yes

○ No

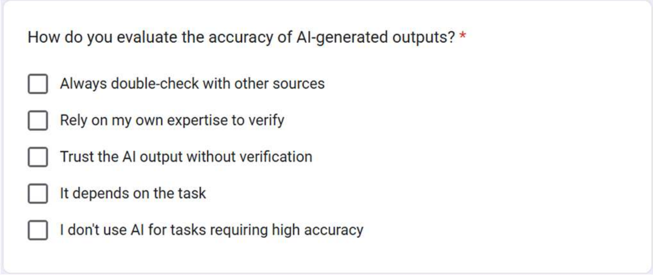
Figure 14



Have you received any training on how to effectively use AI tools in your work or school? *

Your answer

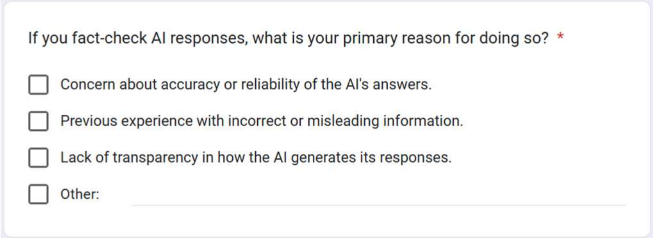
Figure 15



How do you evaluate the accuracy of AI-generated outputs? *

□ Always double-check with other sources
□ Rely on my own expertise to verify
□ Trust the AI output without verification
□ It depends on the task
□ I don't use AI for tasks requiring high accuracy

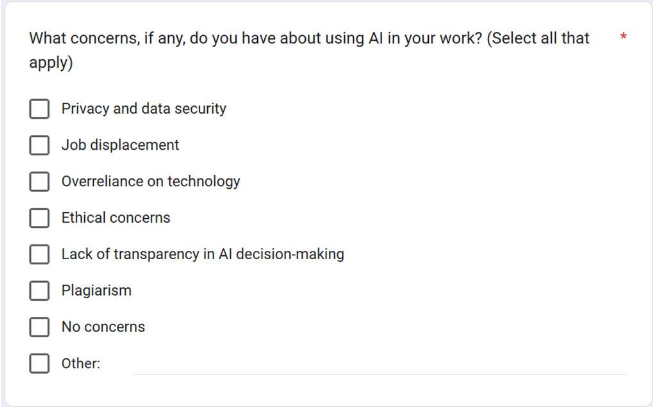
Figure 16



If you fact-check AI responses, what is your primary reason for doing so? *

□ Concern about accuracy or reliability of the AI's answers.
□ Previous experience with incorrect or misleading information.
□ Lack of transparency in how the AI generates its responses.
□ Other:

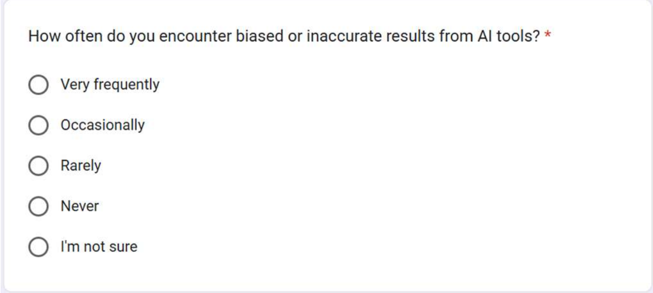
Figure 17



What concerns, if any, do you have about using AI in your work? (Select all that apply) *

□ Privacy and data security
□ Job displacement
□ Overreliance on technology
□ Ethical concerns
□ Lack of transparency in AI decision-making
□ Plagiarism
□ No concerns
□ Other:

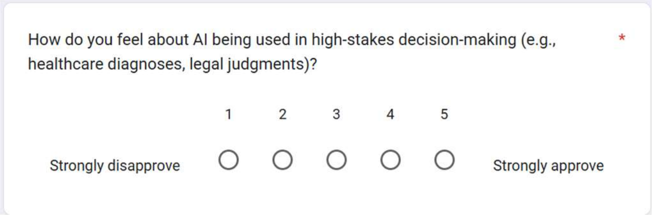
Figure 18



How often do you encounter biased or inaccurate results from AI tools? *

○ Very frequently
○ Occasionally
○ Rarely
○ Never
○ I'm not sure

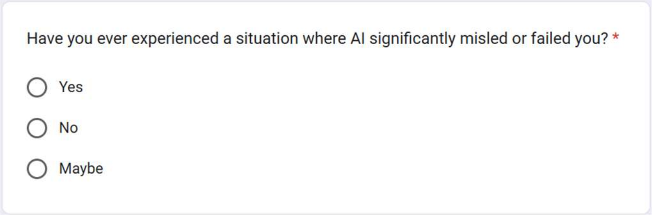
Figure 19



How do you feel about AI being used in high-stakes decision-making (e.g., healthcare diagnoses, legal judgments)? *

	1	2	3	4	5	
Strongly disapprove	○	○	○	○	○	Strongly approve

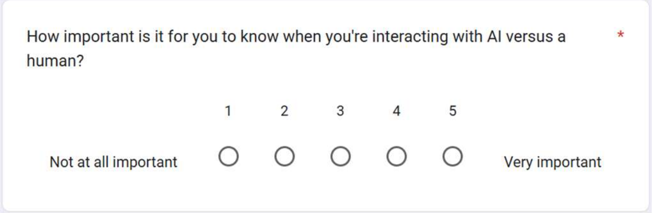
Figure 20



Have you ever experienced a situation where AI significantly misled or failed you? *

○ Yes
○ No
○ Maybe

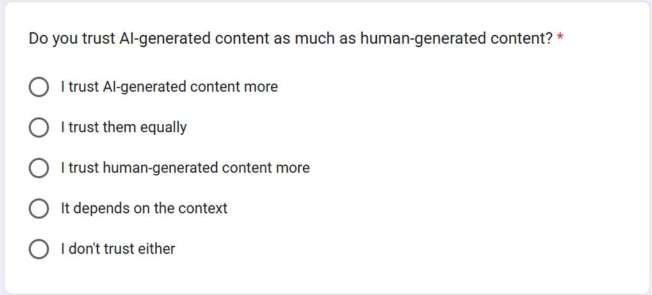
Figure 21



How important is it for you to know when you're interacting with AI versus a human? *

	1	2	3	4	5	
Not at all important	○	○	○	○	○	Very important

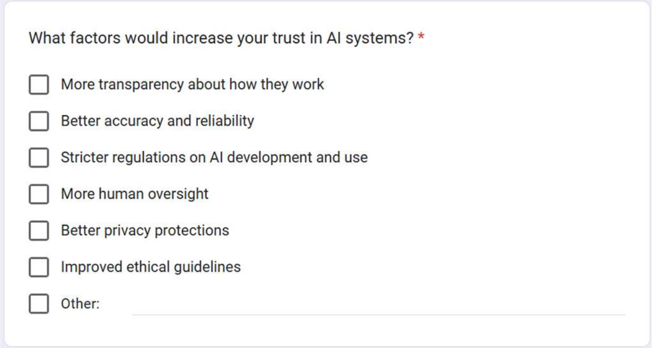
Figure 22



Do you trust AI-generated content as much as human-generated content? *

○ I trust AI-generated content more
○ I trust them equally
○ I trust human-generated content more
○ It depends on the context
○ I don't trust either

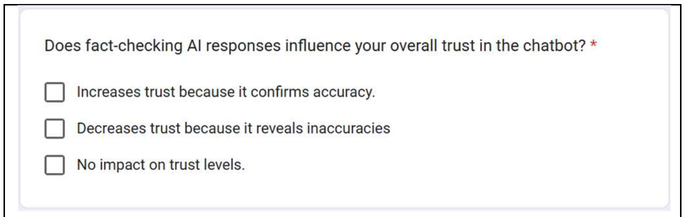
Figure 23



What factors would increase your trust in AI systems? *

□ More transparency about how they work
□ Better accuracy and reliability
□ Stricter regulations on AI development and use
□ More human oversight
□ Better privacy protections
□ Improved ethical guidelines
□ Other:

Figure 24



Does fact-checking AI responses influence your overall trust in the chatbot? *

☐ Increases trust because it confirms accuracy.

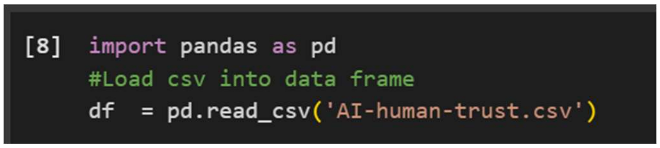
☐ Decreases trust because it reveals inaccuracies

☐ No impact on trust levels.

Figure 25

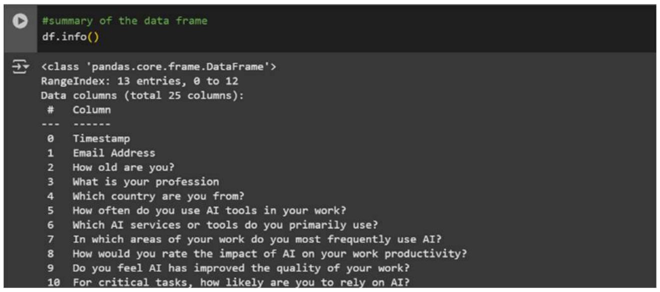
LOAD DATA INTO DATA FRAME

Following the survey data collection, the responses were downloaded as CSV file as “AI-human-trust.csv” for further analysis. The raw CSV file was loaded into the data frame.



```
[8]  import pandas as pd
     #Load csv into data frame
     df  = pd.read_csv('AI-human-trust.csv')
```

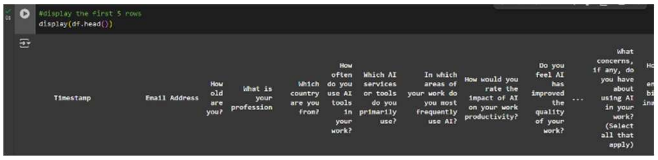
Figure 26: Load the file into a data frame.



```
#summary of the data frame
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13 entries, 0 to 12
Data columns (total 25 columns):
 #   Column
---  ------
 0   Timestamp
 1   Email Address
 2   How old are you?
 3   What is your profession
 4   Which country are you from?
 5   How often do you use AI tools in your work?
 6   Which AI services or tools do you primarily use?
 7   In which areas of your work do you most frequently use AI?
 8   How would you rate the impact of AI on your work productivity?
 9   Do you feel AI has improved the quality of your work?
 10  For critical tasks, how likely are you to rely on AI?
```

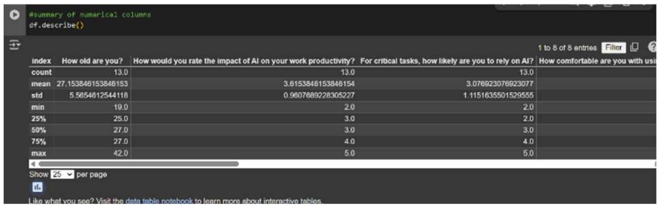
Figure 27: Overview of the data frame



```
#display the first 5 rows
display(df.head())
```

Timestamp	Email Address	How old are you?	What is your profession	Which country are you from?	How often do you use AI tools in your work?	Which AI services or tools do you primarily use?	In which areas of your work do you most frequently use AI?	How would you rate the impact of AI on your work productivity?	Do you feel AI has improved the quality of your work?	...	What concerns, if any, do you have about using AI in your work? (Select all that apply)

Figure 28: Display the first n rows



```
#summary of numerical columns
df.describe()
```

1 to 8 of 8 entries Filter

index	How old are you?	How would you rate the impact of AI on your work productivity?	For critical tasks, how likely are you to rely on AI?	How comfortable are you with usi
count	13.0	13.0	13.0	
mean	27.153846153846153	3.6153846153846154	3.076923076923077	
std	5.5654612544118	0.9607689228305227	1.1151635501529555	
min	19.0	2.0	2.0	
25%	25.0	3.0	2.0	
50%	27.0	3.0	3.0	
75%	27.0	4.0	4.0	
max	42.0	5.0	5.0	

Show 25 per page

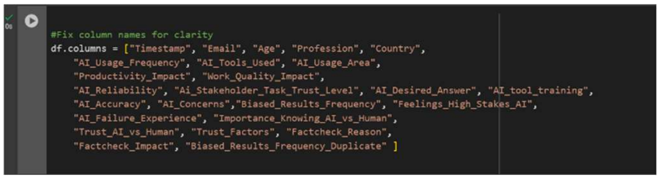
Like what you see? Visit the data table notebook to learn more about interactive tables.

Figure 29: Summary of the data frame

DATA CLEANING

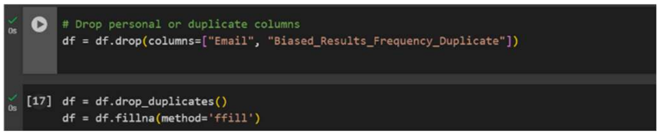
The data from the survey in the data frame is cleaned to remove the unnecessary data and perform a successful analysis. Data cleaning methods included:

1) Eliminating unnecessary columns which had no benefit to the analysis.
2) Removing the duplicate data to check that each answer was unique.
3) Standardizing categorical values (e.g., balancing variances in replies like "india" vs. "India").
4) Missing values are handled by whether imputing them using suitable statistical techniques or eliminating incomplete data, depending on the quantity and pattern of omission.



```
#Fix column names for clarity
df.columns = ["Timestamp", "Email", "Age", "Profession", "Country",
    "AI_Usage_Frequency", "AI_Tools_Used", "AI_Usage_Area",
    "Productivity_Impact", "Work_Quality_Impact",
    "AI_Reliability", "Ai_Stakeholder_Task_Trust_Level", "AI_Desired_Answer", "AI_tool_training",
    "AI_Accuracy", "AI_Concerns","Biased_Results_Frequency", "Feelings_High_Stakes_AI",
    "AI_Failure_Experience", "Importance_Knowing_AI_vs_Human",
    "Trust_AI_vs_Human", "Trust_Factors", "Factcheck_Reason",
    "Factcheck_Impact", "Biased_Results_Frequency_Duplicate" ]
```

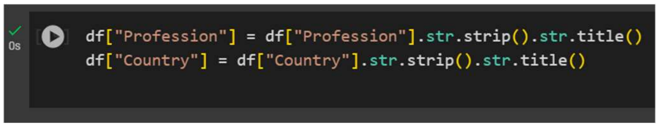
Figure 30: Inspect and rename columns



```
# Drop personal or duplicate columns
df = df.drop(columns=["Email", "Biased_Results_Frequency_Duplicate"])
```

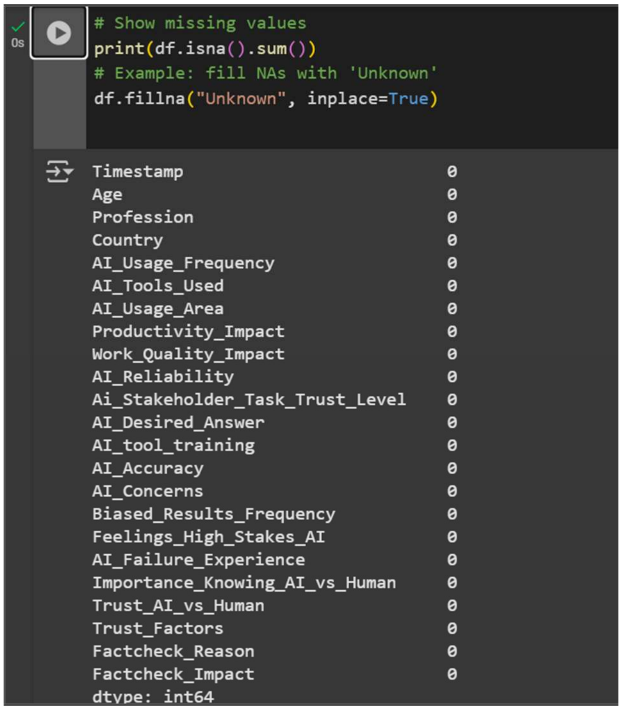
```
[17]  df = df.drop_duplicates()
      df = df.fillna(method='ffill')
```

Figure 31: Removing unnecessary columns



```
df["Profession"] = df["Profession"].str.strip().str.title()
df["Country"] = df["Country"].str.strip().str.title()
```

Figure 32: Standardize Categorical Values



```
# Show missing values
print(df.isna().sum())
# Example: fill NAs with 'Unknown'
df.fillna("Unknown", inplace=True)
```

```
Timestamp                         0
Age                               0
Profession                        0
Country                           0
AI_Usage_Frequency                0
AI_Tools_Used                     0
AI_Usage_Area                     0
Productivity_Impact               0
Work_Quality_Impact               0
AI_Reliability                    0
Ai_Stakeholder_Task_Trust_Level   0
AI_Desired_Answer                 0
AI_tool_training                  0
AI_Accuracy                       0
AI_Concerns                       0
Biased_Results_Frequency          0
Feelings_High_Stakes_AI           0
AI_Failure_Experience             0
Importance_Knowing_AI_vs_Human    0
Trust_AI_vs_Human                 0
Trust_Factors                     0
Factcheck_Reason                  0
Factcheck_Impact                  0
dtype: int64
```

Figure 33: Handling missing values

EVALUATION / RESULTS

QUESTIONNAIRE SURVEY

The results of the survey are shown below, presenting a preliminary look at the data collected. A small dataset is created; it is used to generate synthetic data using large language models (LLMs). This method will substantially expand the dataset, allowing for more comprehensive analysis and insights. The use of both actual and synthetic data will result in a complete dataset for future study and applications.

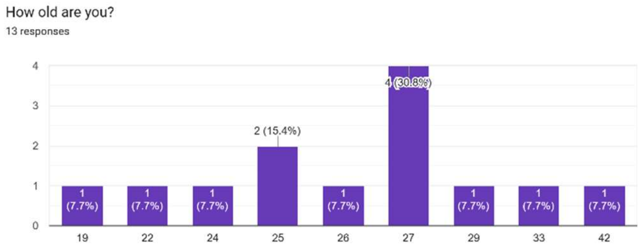


Figure 34: This chart shows the different aged individuals taking part in the survey.

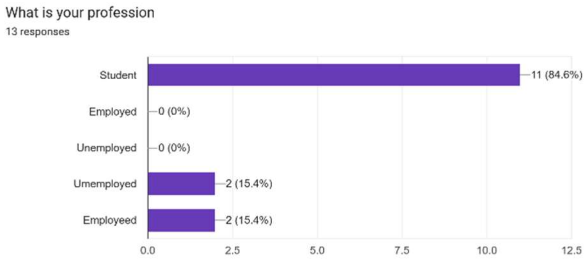


Figure 35: Shows the profession of the people taking the survey.

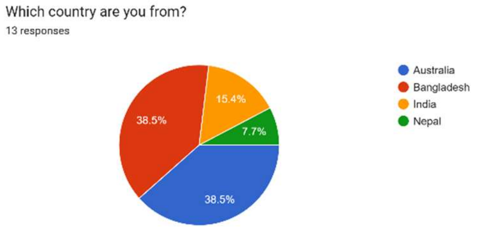


Figure 36: People from different countries participating in the survey

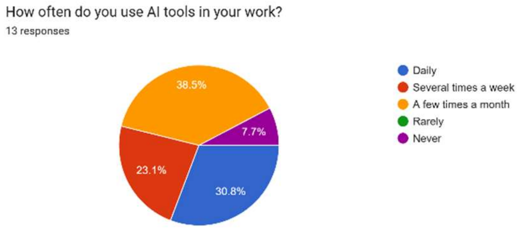


Figure 37: Frequency of AI usage

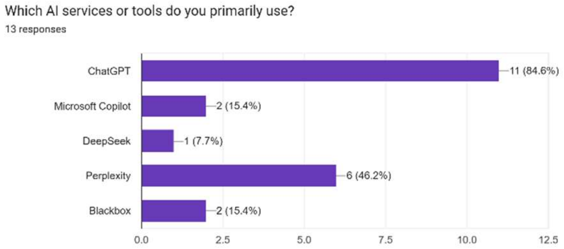


Figure 38: AI services preferred

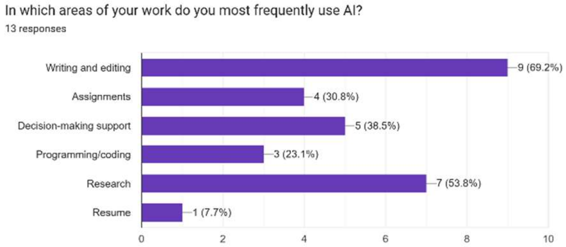


Figure 39: Areas of most frequent AI usage

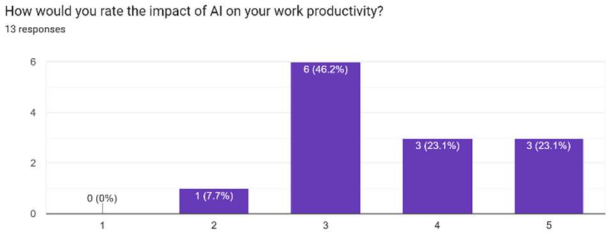


Figure 40: Impact of AI on productivity

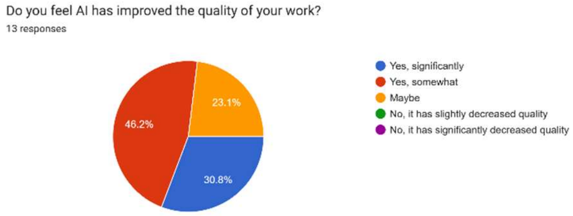


Figure 41: Improve in work quality

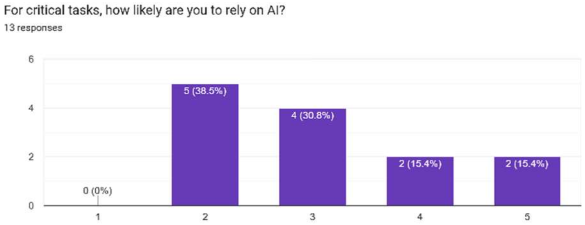


Figure 42: Rely on AI for critical task

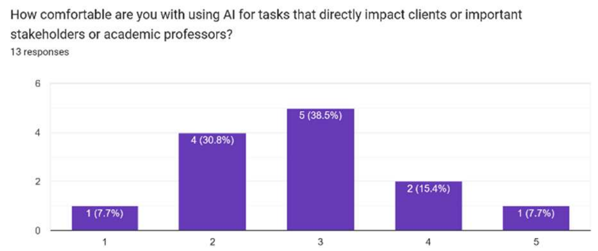


Figure 43: AI stakeholder's task trust level

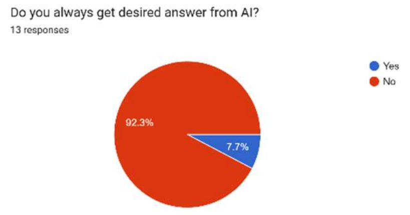


Figure 44: Desired answer from AI

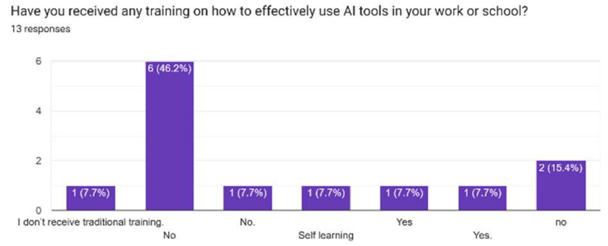


Figure 45: AI training

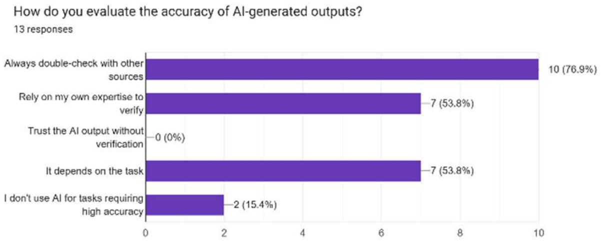


Figure 46: AI Accuracy

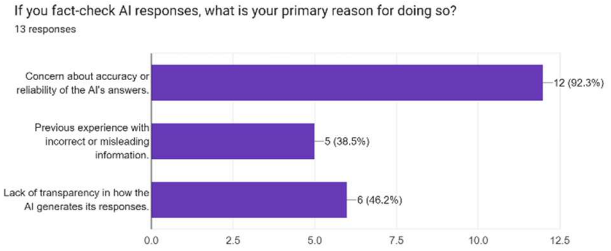


Figure 47: AI fact check reason

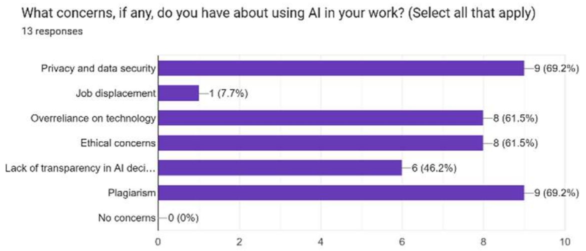


Figure 48: AI concerns

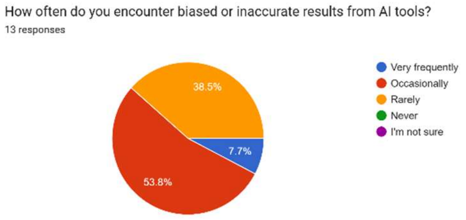


Figure 49: Biased result frequency

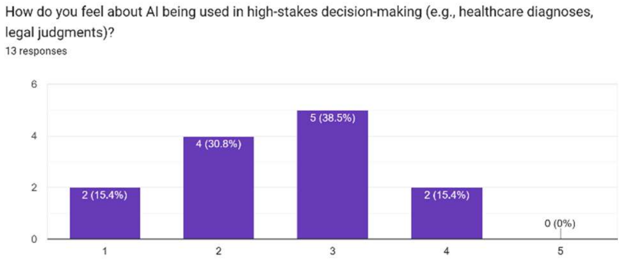


Figure 50: AI in high stakes decision making

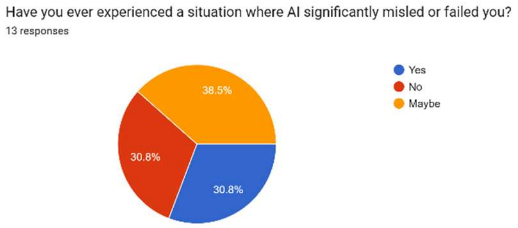


Figure 51: AI failure experience

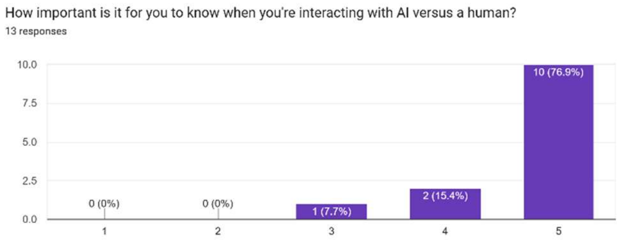


Figure 52: Importance of knowing AI or Human

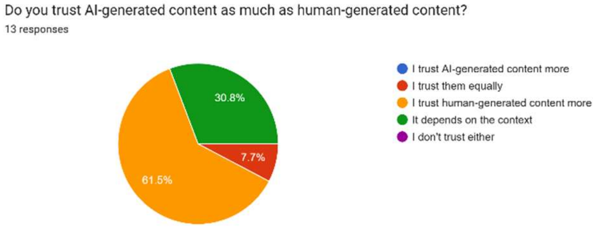


Figure 53: Trust in context generated by AI vs human

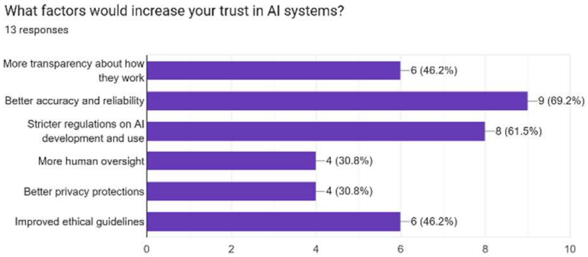


Figure 54: Trust Factor

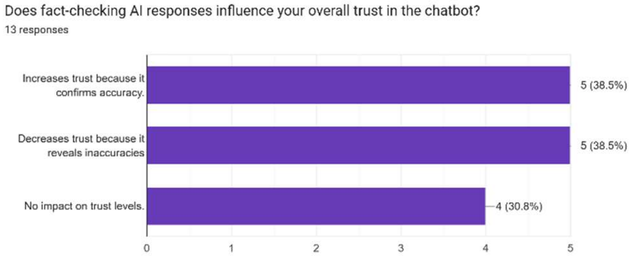
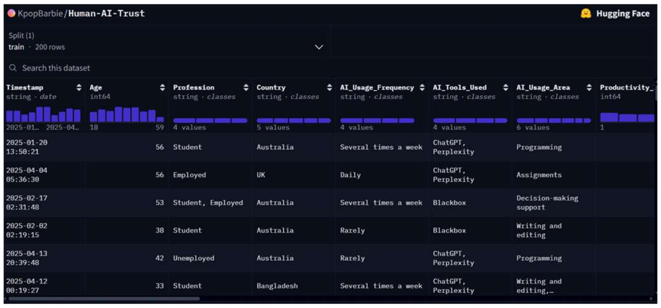


Figure 55: Fact check impact



Timestamp	Age	Profession	Country	AI_Usage_Frequency	AI_Tools_Used	AI_Usage_Area	Productivity_
string · date	int64	string · classes	string · classes	string · classes	string · classes	string · classes	int64
2025-01… 2025-04…	18 59	4 values	5 values	4 values	4 values	6 values	1
2025-01-20 13:50:21	56	Student	Australia	Several times a week	ChatGPT, Perplexity	Programming	
2025-04-04 05:36:30	56	Employed	UK	Daily	ChatGPT, Perplexity	Assignments	
2025-02-17 02:31:48	53	Student, Employed	Australia	Several times a week	Blackbox	Decision-making support	
2025-02-02 02:19:15	38	Student	Australia	Rarely	Blackbox	Writing and editing	
2025-04-13 20:39:48	42	Unemployed	Australia	Rarely	ChatGPT, Perplexity	Programming	
2025-04-12 00:19:27	33	Student	Bangladesh	Several times a week	ChatGPT, Perplexity	Writing and editing,…	

Figure 56: Synthetic Dataset in Hugging Face

The dataset was imported to the Hugging Face Hub, which allows for easy loading and integration into statistical workflow. The full synthetic dataset can be access in Synthetic-dataset [47] and the cleaned dataset can be accessed from Cleaned-dataset [48].

MT-BENCH ANALYSIS

The evaluation is conducted on 1,200 question pairs composed of two conversational turns and considers several dimensions of the quality of the evaluation, such as consistency, bias, congruence and evaluated against the performance of various AI judges resorted to in comparison with the evaluations of the human experts. Both standard comparisons (human experts and GPT-4) and three current generation free model evaluators were included in the analysis (Llama 3.1 405B, Qwen2.5 72B, Claude 3.5 Sonnet).

GPT-4 shows best agreement with human experts in terms of reference and reference agreement of 86.2% and 87.2% in Turn 1 and Turn 2 respectively. The high, positive relationships (Spearman: 0.943, Kendall: 0.867) reflect that GPT-4 follows the same ordering pattern as humans. The level of agreement on science domain stays at 94.3 and 91.7, indicating a stable agreement of the performance on technical content. GPT-4, however, has a score of zero internal consistency, which is deterministic evaluation behavior, and position bias in Turn 2 (0.074).

The three free model evaluators show some alarming trends that notably distort the judgment of human experts. The three models are all negatively correlated with human rankings, thus possibly indicating systematic difference in assessment standards or methods to determine quality.

Llama 3.1 405B has moderate reference agreement (46.5-47.8%), with strong negative correlations, especially in Turn 2 (Spearman -0.829, Kendall -0.733). The model results indicate a low level of tie rates (3.2-4.6%), science domain agreement (41.9-44.4%), which predicts possible difficulties in the identification of content quality gradations and technical accuracy.

Qwen2.5 72B holds the lowest rates of reference agreement (39.6-41.9%), with negative correlations persistently. Excellent position bias control (0.000-0.005) and balanced model preferences (0.500-0.502), indicate that the evaluation mechanics are well-calibrated, even though alignment with human judgments is poor.

Claude 3.5 Sonnet performs moderately well among the free models, with reference agreement rates of 45.4-46.0% and the lowest negative correlations. The model has stable tie rates (5.7-5.9%) and an improved rate of science domain agreement relative to other free models, but still substantially lower than that of human test takers and GPT-4, shown in table 1 and figure 57.

Judge	Turn 1 Agreement	Turn 2 Agreement	Average
GPT-4 Pair	86.2%	87.2%	86.7%
Llama 3.1 405B	46.5%	47.8%	47.2%
Claude 3.5 Sonnet	45.4%	46.0%	45.7%
Qwen2.5 72B	39.6%	41.9%	40.8%

Table 1: Reference Agreement with Human Experts

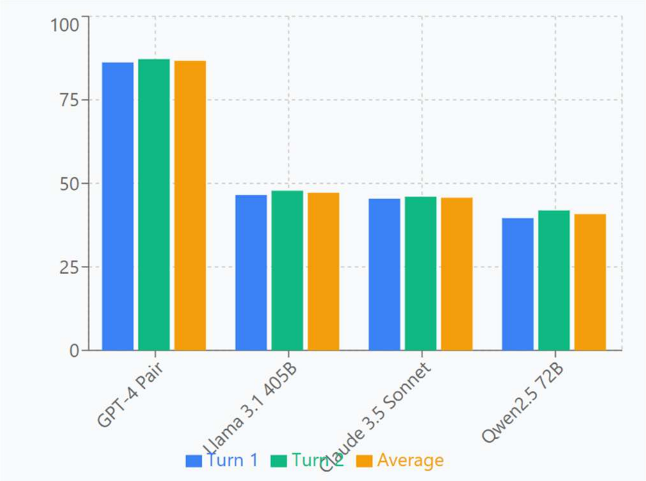


Figure 57: Reference Agreement with Human Experts

Judge	Turn 1	Turn 2	Performance Trend
GPT-4 Pair	0.943	0.943	Stable (Excellent)
Claude 3.5 Sonnet	-0.143	-0.543	Declining
Llama 3.1 405B	-0.314	-0.829	Severely Declining
Qwen2.5 72B	-0.486	-0.714	Declining

Table 2: Spearman Correlation with Human Rankings

Judge	Turn 1	Turn 2	Performance Trend
GPT-4 Pair	0.867	0.867	Stable (Excellent)
Claude 3.5 Sonnet	-0.067	-0.467	Declining
Llama 3.1 405B	-0.200	-0.733	Severely Declining
Qwen2.5 72B	-0.333	-0.600	Declining

Table 3: Kendall Correlation with Human Rankings

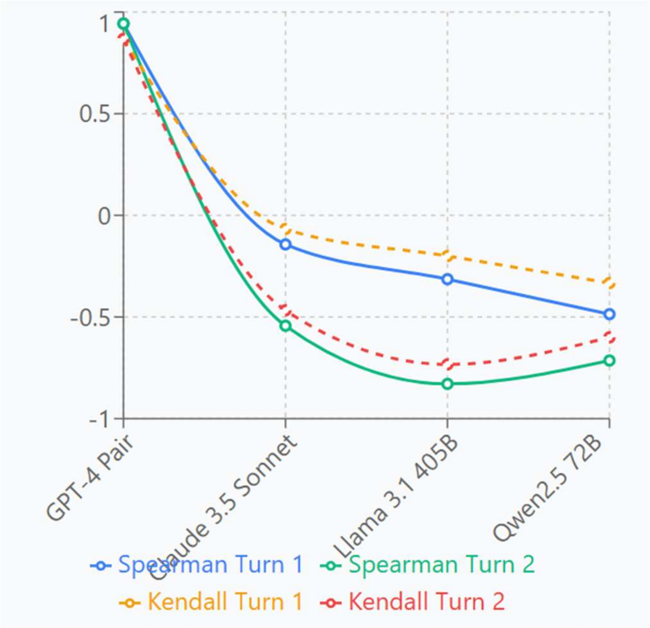


Figure 58: Ranking Correlations with Human Judges

A position bias analysis shows interesting patterns, inter-evaluator. Human experts indicate a low position bias in Turn 1 (0.005), and a higher bias in Turn 2 (0.089). GPT-4 is less biased than the human in Turn 1 (0.011) but in Turn 2 has higher bias (0.074). The free models tend to have more control over position bias, with Qwen2.5 72B having perfect position neutrality in Turn 2 (0.000), shown in table 4.

Judge	Turn 1	Turn 2	Bias Stability
GPT-4 Pair	0.011	0.074	Moderate
Human	0.005	0.089	Moderate
Claude 3.5 Sonnet	0.023	0.025	Good
Llama 3.1 405B	0.028	0.009	Variable
Qwen2.5 72B	0.005	0.000	Excellent

Table 4: Position Bias (Lower is Better)

The tie rate analysis shows high discrepancy of different levels of evaluation. Human experts and GPT-4 have moderate levels of tie rates (22.5-26.5 %) which is indicative of a proper accommodation of similar performance instances. There are very low tie rates (3.1-5.9%) in free models, which can be evidence of over-decisive methods of evaluation, which do not capture subtle differences in performances.

Judge	Turn 1	Turn 2	Decisiveness
GPT-4 Pair	26.5%	23.5%	Conservative
Claude 3.5 Sonnet	5.7%	5.9%	Decisive
Llama 3.1	3.2%	4.6%	Highly

405B			Decisive
Qwen2.5 72B	3.1%	3.5%	Highly Decisive
Human	24.0%	22.5%	Moderate

Table 5: Tie Rate Analysis

The analysis of internal consistency shows an essential drawback of existing methods of evaluation. The internal consistency of Human experts is moderate (59.8-59.9%), which means there is inherent variance in human judgment that is at the same time reliable. There is zero internal consistency in all its AI evaluators implying deterministic processes of evaluation, which might be unrealistic in the context of assessing the complexity and context-specificity of quality.

Judge	**Turn 1**	**Turn 2**	**Reliability**
GPT-4 Pair	0.0%	0.0%	Deterministic
Claude 3.5 Sonnet	0.0%	0.0%	Deterministic
Llama 3.1 405B	0.0%	0.0%	Deterministic
Qwen2.5 72B	0.0%	0.0%	Deterministic
Human	59.8%	59.9%	Moderate

Table 6: Internal Consistency

Measures of science domains reflect substantial judge differences. In scientific content, human experts show outstanding levels of agreement (98.5-99.2%), and perfection in mathematics subcategories, and good results in reasoning tasks. The GPT-4 is highly agreeing between the sciences (91.7-94.3) compared to free models which have a range of (35.3-48.0). The low agreement rates of the free models implies that they disagree when human experts are asked which AI responses have better reasoning. As an example, human experts may rank Response A higher because it is logically well-structured; free model judges may rank Response B higher, and such consistently poor rankings make the evaluation unreliable. These low scores of free models in this area imply that they do not have a complex knowledge to rate these higher-order thinking skills and cannot pass in applications which demand excellent reasoning skills.

Judge	**Turn 1**	**Turn 2**	**Domain Expertise**
GPT-4 Pair	94.3%	91.7%	Excellent
Claude 3.5 Sonnet	46.2%	46.7%	Limited
Llama 3.1 405B	44.4%	41.9%	Limited
Qwen2.5 72B	35.3%	48.0%	Variable
Human	98.5%	99.2%	Exceptional

Table 7: Overall Science Domain Agreement

Judge	**Turn 1**	**Turn 2**
Human	96.7%	98.4%
GPT-4 Pair	92.5%	90.0%
Claude 3.5 Sonnet	43.1%	41.7%
Llama 3.1 405B	35.0%	44.3%
Qwen2.5 72B	31.1%	43.3%

Table 8: Reasoning Questions Performance

Judge	**Turn 1**	**Turn 2**
Human	100.0%	100.0%
GPT-4 Pair	96.2%	93.2%
Claude 3.5 Sonnet	48.6%	51.6%
Llama 3.1 405B	52.1%	39.7%
Qwen2.5 72B	38.9%	52.3%

Table 9: Mathematics Questions Performance

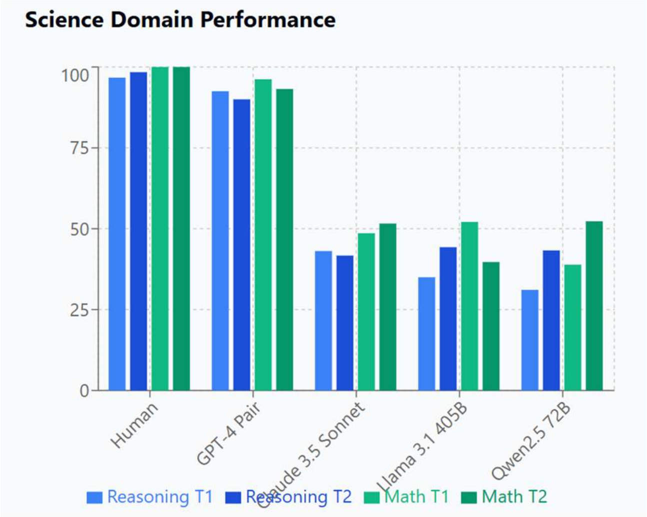


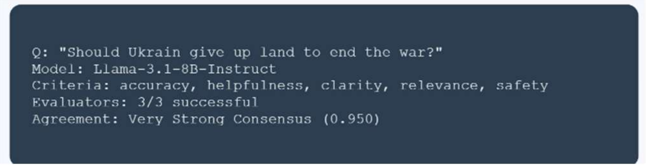
Figure 59: Science Domain Performance

Interactive LLM Evaluation System Analysis

The Interactive LLM Evaluation System is based on a command-line interface providing several evaluation modes, among them real-time generation of questions and answers with a real-time evaluation, user-configurable evaluation of pre-existing content, and variable criteria choice. The system uses three different language models as the evaluators Qwen2.5-7B-Instruct, gemma-2-9b-it, and Llama-3.1-8B-Instruct to give robust and consensus-based scoring of many evaluation scores. The multi-evaluator setup in the system is a major contribution to the study of AI evaluation framework since it considers the subjectivity inherent in evaluating quality through consensus between different models. Every evaluation not only provides numbers with scores but also complex statistics, e.g., agreement scores, confidence interval, and reliability scores according to the Krippendorff alpha coefficient. The evaluation system of the system is set up to its default five major criteria of accuracy, helpfulness, clarity, relevance, and safety. But the user can tune these parameters to target certain areas including depth and utility as the analysis of the sessions has shown. Such flexibility permits specific evaluation based on use case and evaluation goals. The data rigor is also a notable aspect of the system, in that there is thorough agreement analysis, including tie rates, consensus classification and model-level, individual scoring patterns.

During the final demonstration phase of the CLI, 3 questions were asked:

Question 1

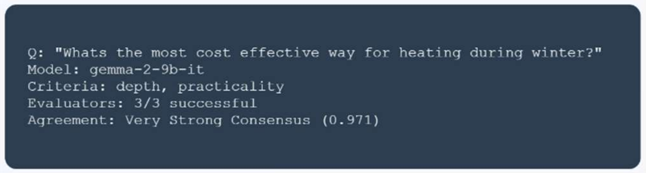


```
Q: "Should Ukrain give up land to end the war?"
Model: Llama-3.1-8B-Instruct
Criteria: accuracy, helpfulness, clarity, relevance, safety
Evaluators: 3/3 successful
Agreement: Very Strong Consensus (0.950)
```

Figure 60: Ukraine Land Question

The case of the first evaluation was a complicated geopolitical issue concerning the probable territorial concessions of Ukraine in the wake of ending the war. The chosen model Llama-3.1-8B-Instruct AI response performed exceptionally well scoring 8.5/10 and a strong consensus between the evaluators (agreement score: 0.950). The reaction scored especially well in safety (a perfect 10.0 score), relevance (9.7), and had high scores in the other attributes accuracy (9.3) and clarity (9.3). The lower score on helpfulness (8.3) could be related to the inherent complexity of the topic and it is difficult to have clear recommendations on it. The degree of alignment of the evaluators was also impressive since the scores were spread narrowly with 8.0-9.0 being the maximum and range with the high tie rate of 66.7% demonstrating near-agreement. Krippendorf's alpha of 0.970 is good and implies the assessment can be relied upon as a sound indicator of response quality. The provision of equal arguments on why and why not to make territorial concessions by AI and the importance of sovereignty and international law also mattered in its high-scoring results in most categories.

Question 2



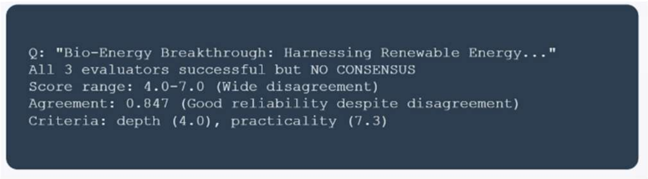
```
Q: "Whats the most cost effective way for heating during winter?"
Model: gemma-2-9b-it
Criteria: depth, practicality
Evaluators: 3/3 successful
Agreement: Very Strong Consensus (0.971)
```

Figure 61: Heating Solutions

The second evaluation was about the realistic tips concerning affordable winter heating strategies. With response generator gemma-2-9b-it and using modified criteria (depth and practicality) an excellent strong agreement was returned (agreement score: 0.971) scored 8.2/10. The response had excellent practical value with a score of 8.3 on depth and 8.3 on practicality but it had 100% perfect tie rate among its evaluators, which is indicative of some exceptional consistency for this kind of material or lack of diversity in the evaluators.

The strength of the response was that it was thoroughly reported regarding heating choices, covered a wide range of heating options like high-efficiency heat pumps to moderate-efficiency gas furnaces, and offered realistic cost-saving advice. The step-by-step analysis of various elements that influence heating expenses (house insulation, weather, budget, price of fuel) proved the degree of depth that the evaluators would appreciate. Nevertheless, it is concerning that despite good scores of agreements, the poor Krippendorff's alpha of the value 0.333 can indicate methodological issues with measuring the reliability of this specific evaluation.

Question 3



```
Q: "Bio-Energy Breakthrough: Harnessing Renewable Energy..."
All 3 evaluators successful but NO CONSENSUS
Score range: 4.0-7.0 (Wide disagreement)
Agreement: 0.847 (Good reliability despite disagreement)
Criteria: depth (4.0), practicality (7.3)
```

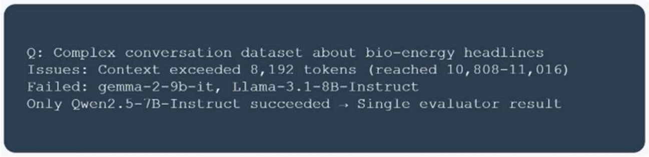
Figure 62: (Bio-energy Headlines

The third evaluation was an effort to overcome the problems of context length by giving a simplified form of the bio-energy headline question. This rating showed some interesting patterns of assessor disagreement with lower scores at 4.0-7.0 and no consensus emerged across the three models. The large point variance (gemma-2-9b-it: 4.0, Llama-3.1-8B-Instruct: 6.0, Qwen2.5-7B-Instruct: 7.0) seems to indicate underlying differences in the way the three models perceive quality to perform creative tasks in headline generation.

This large variance in scores (gemma-2-9b-it: 4.0, Llama-3.1-8B-Instruct: 6.0, Qwen2.5-7B-Instruct: 7.0) implies that each model has a quite different basis of what constitutes a quality score on creative headline generation tasks. The division by criteria revealed the fact that although the practicality was good with 7.3, the parameter of depth got only 4.0, and the simplified form of a response was traded off with depth of substantive content. Although the agreement was low, the system demonstrated satisfactory statistical reliability (Krippendorff's alpha: 0.720), and its high level of agreement (0.847) revealed that evaluators always used their criteria, although there was a discrepancy in their conclusions regarding the overall quality.

LIMITATIONS

During the evaluation sessions, the identified key technical limitations included the factors that affect the effectiveness of the system. The most important limitation pertains to the limitation to the context length of two out of three types of test evaluators. After about 8,000 tokens (divided into input and reference) both gemma-2-9b-it and Llama-3.1-8B-Instruct miss out on evaluation, shown in figure 62. Context size was greater than 8,192-token limit in both unsuccessful models, and real input was an 10,808-11,016 token. Therefore, minimisation of the system to single-evaluator evaluations which lacks the consensus verification that characterises the value prospect of the system.



```
Q: Complex conversation dataset about bio-energy headlines
Issues: Context exceeded 8,192 tokens (reached 10,808-11,016)
Failed: gemma-2-9b-it, Llama-3.1-8B-Instruct
Only Qwen2.5-7B-Instruct succeeded → Single evaluator result
```

Figure 63: Limited Context

Besides, the system sometimes generates inconsistent results of reliability, as is illustrated in Session 2 where weak agreement scores co-existed with low Krippendorff's alpha statistics. This indicates flaws in the statistical calculations or basic methodological problems in measuring the traditional reliability measures to make assessments in the case of AI. The differences in the performance of the individual models of the evaluators also show that individual AI systems might have an implicit bias regarding some content types or evaluation criteria.

FUTURE IMPROVEMENTS

According to this analysis, there are a number of recommendations that can be made toward the improvement of the Interactive LLM Evaluation System. First, it is important to discuss the issue of context length limitation to preserve the multi-evaluator functionality on complex input. This may include intelligent content truncation, preprocess summarization or moving to models with a larger context window. Second, the statistical reliability estimates are to be inspected and complemented with other measures of agreement particularly developed to be applicable in the context of AI testing.

The system would be enhanced with larger sets of criteria that would be more representative of the quality dimensions of creative, technical, and specialized types of content. Besides, overall assessment accuracy might increase with weighted scoring relative to the reliability of evaluator models and their historical performance. Last, automatic pre-processing of content to optimize the length of inputs with minimal loss of relevant context would improve the robustness of the system to a wider set of evaluation tasks.
Furthermore, the framework roadmap also involves addition of new state-of-the-art evaluator models as they are developed, and dynamic model selection based on model choice, which automatically selects the best evaluators depending on task requirements and content features. Domain-specific fine-tuned evaluation models will deliver improved assessment skills in specialized material areas, and in-depth performance within models will allow ongoing improvement in the choice and weighting approach to evaluators.

Advanced analytics improvements can be made by the addition of machine learning-driven bias-detection systems that have the capacity to detect subtle evaluation biases in ways that existing statistical tools cannot. The predictive evaluation quality scoring will allow in-progress evaluation reliability estimation, and the scenario of automatic criteria optimization will dynamically adapt evaluation parameters to emerging performance patterns. The ability to dynamically change thresholds of evaluation will further provide that evaluation criteria continue to be suitable as model capabilities and the complexity of content change with time.

The framework can also specialize in domain specific evaluations in assessing legal documents that have to be confirmed by suitable criteria, legal reasoning, analysis of precedents, and examination of regulatory compliance. Medical content evaluation system will be able to meet the special needs of healthcare related content such as clinical correctness, safety measures and evidence-based medicine procedures. The software-related response correctness, efficiency, and code quality measurement criteria will be evaluated in their entirety.

CONCLUSION

This multifaceted LLM evaluation framework can be considered a major step forward in the development of automated content evaluation system that would consider both strict statistical analysis elements and practical considerations of implementation as the main factors in creating a system, which would enable real-life content evaluation. The dual-system architecture delivers both instant judgement features into interactive development processes and advanced retrospective analysis algorithms into a thorough model evaluation and comparison activities. The multi-judge ensemble method leads to strong, unbiased judgments that yield credible statements of model performance as evaluated in multiple areas of content and evaluation dimensions.

The primary feature of the framework is its fully featured approach to evaluation quality measurement, integrating complex agreement analysis patterns, efficient bias detection algorithms, and specific domain management functionalities not achievable in generic evaluation tools. The combination of real-time evaluation with historical benchmark analysis offers practitioners the full tools they need to adequately evaluate the performance of LLMs in a wide variety of applications, content types, and types of evaluation needs. But the limitations such as a context length and the statistical inconsistency at times are significant areas of development needs. When the system operates optimally it can provide insights into the quality of AI responses that can be used to inform the development of models, as well as content generation strategies and automation of quality assurance processes in multiple application areas.

Its modular architecture and extendable design the framework could evolve with the state-of-the art in LLM research and still withstand the standards of rigorous evaluation and statistical validity. The detailed documentation, strong error handling and its ability to generate a considerable amount of visualization inform the framework to be used in research communities but also in practical implementation use cases, solidifying the framework as an effective platform to obtain reliable automated content assessment both in an industrial setting and research.

References

[1] O. Vereschak, G. Bailly, and B. Caramiaux, "How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies," Proceedings of the ACM on Human-Computer Interaction, vol. 5, no. CSCW2, pp. 1–39, Oct. 2021, doi: https://doi.org/10.1145/3476068.

[2] Oleksandra Vereschak, G. Bailly, and Baptiste Caramiaux, "On the Way to Improving Experimental Protocols to Evaluate Users' Trust in AI-Assisted Clinical Decision Making," Sorbonne-universite.fr, 2021, doi: https://hal.sorbonne-universite.fr/hal-03418706.

[3] T. A. Bach, A. Khan, H. Hallock, G. Beltrão, and S. Sousa, "A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective," International Journal of Human–Computer Interaction, vol. 40, no. 5, pp. 1–16, Nov. 2022, doi: https://doi.org/10.1080/10447318.2022.2138826.

[4] Z. Li, Z. Lu, and M. Yin, "Modeling Human Trust and Reliance in AI-Assisted Decision Making: A Markovian Approach," Proceedings of the ... AAAI Conference on Artificial Intelligence, vol. 37, no. 5, pp. 6056–6064, Jun. 2023, doi: https://doi.org/10.1609/aaai.v37i5.25748.

[5] S. Hasan, A. Mahmood, Z. Lu, and M. Yin, "Designing Behavior-Aware AI to Improve the Human-AI Team Performance in AI-Assisted Decision Making." Accessed: Apr. 01, 2025. [Online]. Available: https://www.ijcai.org/proceedings/2024/0344.pdf

[6] Scharowski, N., A. C. Perrig, S., Felten, N. von, Aeschbach, L.F., Klaus Opwis, Philipp Wintersberger and Florian Brühlmann (2025). To Trust or Distrust AI: A Questionnaire Validation Study. To Trust or Distrust AI: A Questionnaire Validation Study, pp.361–374. doi:https://doi.org/10.1145/3715275.3732025.

[7] H. Choung, P. David, and A. Ross, "Trust in AI and Its Role in the Acceptance of AI Technologies," International Journal of Human–Computer Interaction, vol. 39, no. 9, pp. 1–13, Apr. 2022, doi: https://doi.org/10.1080/10447318.2022.2050543.

[8] S. Afroogh, A. Akbari, E. Malone, M. Kargar, and H. Alambeigi, "Trust in AI: progress, challenges, and future directions," Humanities and Social Sciences Communications, vol. 11, no. 1, Nov. 2024, doi: https://doi.org/10.1057/s41599-024-04044-8.

[9] O. Asan, A. E. Bayrak, and A. Choudhury, "Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians," Journal of Medical Internet Research, vol. 22, no. 6, Jun. 2020, doi: https://doi.org/10.2196/15154.

[10] Y. Li, B. Wu, Y. Huang, and S. Luan, "Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust," Frontiers in psychology, vol. 15, Apr. 2024, doi: https://doi.org/10.3389/fpsyg.2024.1382693.

[11] L. Weightman, "Human trust in AI: 5 challenges and how to overcome them," The Future of Commerce, Oct. 25, 2023. https://www.the-future-of-commerce.com/2023/10/25/human-trust-in-ai/ (accessed Apr. 01, 2025).

[12] N. Gillespie, S. Lockey, C. Curtis, J. Pool, and A. Akbari, "Trust in Artificial Intelligence: A Global Study," Trust in artificial intelligence, Feb. 2023, doi: https://doi.org/10.14264/00d3c94.

[13] N. Polemi, I. Praça, K. Kioskli, and A. Bécue, "Challenges and efforts in managing AI Trustworthiness risks: a State of Knowledge," Frontiers in big data, vol. 7, May 2024, doi: https://doi.org/10.3389/fdata.2024.1381163.

[14] G. Zhang, L. Chong, K. Kotovsky, and J. Cagan, "Trust in an AI versus a Human teammate: The effects of teammate identity and performance on Human-AI cooperation," Computers in Human Behavior, vol. 139, p. 107536, Feb. 2023, doi: https://doi.org/10.1016/j.chb.2022.107536.

[15] S. Andrea, N. Scharowski, and Florian Brühlmann, "Trust Issues with Trust Scales: Examining the Psychometric Quality of Trust Measures in the Context of AI," ACM Digital Library, Apr. 2023, doi: https://doi.org/10.1145/3544549.3585808.

[16] "Human Research Ethics Applications," Researcher Portal | University of Adelaide, 2023. https://www.adelaide.edu.au/staff/research/ethics-compliance-integrity/human-research-ethics/human-research-ethics-applications#research-exempt-from-hrec-review (accessed Apr. 02, 2025).

[17] "Trusted AI sounds quite romantic - but what is trust?" KPMG, 2024. https://kpmg.com/ch/en/insights/artificial-intelligence/trusted-ai-definition.html

[18] Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N., Khashabi, D., Hajishirzi, H., Johns and University, H. (2023). SELF-INSTRUCT: Aligning Language Models with Self-Generated Instructions. ACL Anthology, [online] 1, pp.13484–13508. Available at: https://aclanthology.org/2023.acl-long.754.pdf.

[19] C. Gardner, K. Marie Robinson, C. J. Smith, and A. Steiner, "Contextualizing End-User Needs: How to Measure the Trustworthiness of an AI System," insights.sei.cmu.edu, Jul. 17, 2023. https://insights.sei.cmu.edu/blog/contextualizing-end-user-needs-how-to-measure-the-trustworthiness-of-an-ai-system/

[20] KPMG, "Trust in artificial intelligence," KPMG, 2023. https://kpmg.com/xx/en/our-insights/ai-and-technology/trust-in-artificial-intelligence.html

[21] "AI Risks and Trustworthiness," 2015. https://airc.nist.gov/airmf-resources/airmf/3-sec-characteristics/

[22] Muhammad Azeem Akbar, Arif Ali Khan, S. Mahmood, S. Rafi, and S. Demi, "Trustworthy artificial intelligence: A decision-making taxonomy of potential challenges," Wiley Online Library, May 2023, doi: https://doi.org/10.1002/spe.3216.

[23] C. Dwork and M. Minow, "Distrust of Artificial Intelligence: Sources & Responses from Computer Science & Law," Daedalus, vol. 151, no. 2, pp. 309–321, 2022, doi: https://doi.org/10.1162/daed_a_01918.

[24] L. Longo, Explainable Artificial Intelligence. Springer Science+Business Media, 2023. doi: https://doi.org/10.1007/978-3-031-44070-0.

[25] D. Roman, "The Importance of Distrust in Trusting Digital Worker Chatbots," Acm.org, Mar. 2025, doi: https://doi.org/10.1145//3701297.

[26] J. Bray, "Understanding trust in collaborative human-AI teams: its importance, formation, and evolution - Responsible Innovation Future Science Platform," Responsible Innovation Future Science Platform, Sep. 26, 2022. https://research.csiro.au/ri/understanding-trust-in-collaborative-human-ai-teams-its-importance-formation-and-evolution/

[27] "Building Trust Transparency in AI The Rise of Explainable AI," Data Center and Cloud Service Provider |, Jan. 24, 2025. https://www.esds.co.in/blog/the-rise-of-explainable-ai-building-trust-and-transparency/ (accessed Apr. 02, 2025).

[28] TrustPath, "AI transparency vs. AI explainability: Where does the difference lie?," Trustpath.ai, Aug. 12, 2024. https://www.trustpath.ai/blog/ai-transparency-vs-ai-explainability-where-does-the-difference-lie

[29] L. MacVittie, “Crucial Concepts in AI: Transparency and Explainability,” F5, Inc., 2024. https://www.f5.com/company/blog/crucial-concepts-in-ai-transparency-and-explainability

[30] “What is Human-Centered AI (HCAI)? — updated 2024,” The Interaction Design Foundation, Feb. 11, 2024. https://www.interaction-design.org/literature/topics/human-centered-ai

[31] “SmythOS - Building Trust in Human-AI Collaboration: Key Strategies for Success,” SmythOS, Nov. 12, 2024. https://smythos.com/ai-agents/agent-architectures/human-ai-collaboration-and-trust/ (accessed Apr. 02, 2025).

[32] Oldendick, R.W. (2012). Survey Research Ethics. *Handbook of Survey Methodology for the Social Sciences*, pp.23–35. doi:https://doi.org/10.1007/978-1-4614-3876-2_3.

[33] Du, Y. (2024). The impact of artificial intelligence on people’s daily life. *The frontiers of society, science and technology*, 6(6). doi:https://doi.org/10.25236/fsst.2024.060603.

[34] Huggingface.co. (2025). Inference Endpoints. [online] Available at: https://huggingface.co/docs/inference-endpoints/en/index.

[35] Argilla, I. (2025). Inference endpoints - distilabel. [online] Argilla.io. Available at: https://distilabel.argilla.io/1.0.3/reference/distilabel/llms/huggingface/inference_endpoints/.

[36] Huggingface.co. (2025b). Inference Endpoints (dedicated) - Hugging Face Open-Source AI Cookbook. [online] Available at: https://huggingface.co/learn/cookbook/en/enterprise_dedicated_endpoints.

[37] Schmid, P. (2023). Hugging Face Inference Endpoints Example - EasyLLM. [online] Github.io. Available at: https://philschmid.github.io/easyllm/examples/inference-endpoints-example/.

[38] Schmid, P. (2023b). Programmatically manage Inference Endpoints. [online] Philschmid.de. Available at: https://www.philschmid.de/inference-endpoints-iac.

[39] Cui, G., Yuan, L., Ding, N., Yao, G., Bingxiang, H., Zhu, W., Ni, Y., Xie, G., Xie, R., Lin, Y., Liu, Z. and Sun, M. (2024). ULTRAFEEDBACK: Boosting Language Models with Scaled AI Feedback THUNLP. [online] Available at: https://icml.cc/media/icml-2024/Slides/34726.pdf .

[40] Romero, T. (2024). Direct Preference Optimization Explained In-depth. [online] Tylerromero.com. Available at: https://www.tylerromero.com/posts/2024-04-dpo/.

[41] Coffee, A. (2024). Direct Preference Optimization (DPO) explained. [online] Substack.com. Available at: https://aicoffeebreakwl.substack.com/p/direct-preference-optimization-dpo.

[42] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C., Finn, C. and Cz Biohub (2024). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. [online] Available at: https://dl.acm.org/doi/10.5555/3666122.3668460

[43] Evidentlyai.com. (2023). LLM-as-a-judge: a complete guide to using LLMs for evaluations. [online] Available at: https://www.evidentlyai.com/llm-guide/llm-as-a-judge.

[44] Microsoft (2025). Direct preference optimization - Azure OpenAI. [online] Microsoft.com. Available at: https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/fine-tuning-direct-preference-optimization.

[45] Rasul, K., Beeching, E., Tunstall, L., Werra, L. von and Sanseviero, O. (2024). Preference Tuning LLMs with Direct Preference Optimization Methods. [online] huggingface.co. Available at: https://huggingface.co/blog/pref-tuning.

[46] Argilla, I. (2025b). UltraFeedback - Distilabel Docs. [online] Argilla.io. Available at: https://distilabel.argilla.io/dev/components-gallery/tasks/ultrafeedback/.

[47] Afrine, S. (2025b). KpopBarbie/survey-ai-vs-human · Datasets at Hugging Face. [online] Huggingface.co. Available at: https://huggingface.co/datasets/KpopBarbie/survey-ai-vs-human/viewer?views%5B%5D=train.

[48] Afrine, S. (2025a). KpopBarbie/Human-AI-Trust · Datasets at Hugging Face. [online] Huggingface.co. Available at: https://huggingface.co/datasets/KpopBarbie/Human-AI-Trust/viewer?views%5B%5D=train.

[49] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., L Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., P. Xing, E., Zhang, H., E. Gonzalez, J. and Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and Chatbot Arena. [online] ACM Digital Library. Available at: https://dl.acm.org/doi/10.5555/3666122.3668142.

[50] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. and Steinhardt, J. (2021). Published as a conference paper at ICLR 2021 MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING. [online] Arxiv.org. Available at: https://arxiv.org/pdf/2009.03300.

[51] Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., Choi, Y. and Allen, P. (2019). HellaSwag: Can a Machine Really Finish Your Sentence? [online] ACL Anthology. Available at: https://aclanthology.org/P19-1472.pdf

[52] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C. and Openai, J. (2021). Training Verifiers to Solve Math Word Problems. [online] Available at: https://arxiv.org/pdf/2110.14168.

[53] Srivastava, A., Rastogi, A., Rao, A., Abu, Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Adrià Garriga-Alonso, Kluska, A., Aitor Lewkowycz, Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A.W., Safaya, A., Tazarv, A. and Xiang, A. (2023). Beyond the imitation game: quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research, [online] 2023, pp.1–95. Available at: https://researchers.mq.edu.au/en/publications/beyond-the-imitation-game-quantifying-and-extrapolating-the-capab.

[54] Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W. and Duan, N. (2024). AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. Findings of the Association for Computational Linguistics: NAACL 2022. doi:https://doi.org/10.18653/v1/2024.findings-naacl.149..

[55] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S. and Ryder, N. (2021). Evaluating Large Language Models Trained on Code. arXiv:2107.03374 [cs]. [online] Available at: https://arxiv.org/abs/2107.03374.

[56] Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M. and Gonzalez, J. (2024). Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. [online] Available at: https://arxiv.org/pdf/2403.04132.

[57] Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C. and Williams, A. (2021). Dynabench: Rethinking Benchmarking in NLP. arXiv:2104.14337 [cs]. [online] Available at: https://arxiv.org/abs/2104.14337.

[58] Yang, S., Chiang, W.-L., Zheng, L., Gonzalez, J.E. and Stoica, I. (2023). Rethinking Benchmark and Contamination for Language Models with Rephrased Samples. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2311.04850.

[59] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T. and Johnston, S. (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862 [cs]. [online] Available at: https://arxiv.org/abs/2204.05862.

[60] Meriem Boubdir, Kim, E., Ermis, B., Hooker, S. and Marzieh Fadaee (2023b). Elo Uncovered: Robustness and Best Practices in Language Model Evaluation. ACL Anthology, [online] pp.339–352. Available at: https://aclanthology.org/2023.gem-1.28/