Advanced Natural Learning Processing

Internal Project

Submitted to
School of Technology,
Woxsen University, Hyderabad

Submitted

To

Dr. Uday Chandra

Professor

for

School of Technology

Submitted By

Himanshu Maurya (2068), Rishita Lakshmi (2059),

Disha Agarwal (2052), Aashutosh Gautam (2005)

B.Tech 2023-27

AIML Tigers

February 2026

A Comparative Framework for Evaluating Reasoning, Hallucination, and Safety Alignment in Large Language Models

Abstract

As Large Language Models (LLMs) move from research artefacts to crucial enterprise deployments, it is more important than ever to properly evaluate their reliability, safety, and reasoning capabilities. While architectural developments have improved overall functionality, concerns such as semantic hallucination, representational bias, and misalignment persist. This study proposes a thorough, automated assessment approach for comparing the evolution of LLMs across three architectural paradigms: a basic decoder (GPT-2), an instruction-tuned encoder-decoder (FLAN-T5-Large), and an RLHF-aligned model (LLaMA-3-8B-Instruct). We create and use an 800-prompt, multi-domain dataset that includes factual memory, logical reasoning, ambiguous situations, and ethical problems.Using an automated scoring engine, we quantify model factuality with lexical overlap heuristics, test safety with severity-scaled refusal detection, and assess representational bias with gender pronoun swaps in ambiguous scenarios. Our empirical findings demonstrate the strong evolutionary differences between designs. LLaMA-3-8B displayed superior factual adherence (41% overlap) and strong safety alignment (64% refusal rate on critical violations), effectively resisting contextual jailbreaks. In contrast, FLAN-T5 showed exceptional compliance (0% refusal) but severe hallucinations, whereas GPT-2 failed both safety and coherence restrictions. Furthermore, the study discovers a quantifiable "safety tax," in which enhanced alignment occasionally results in false-positive refusals on harmless enquiries. Finally, this paradigm establishes a replicable mechanism for auditing LLM trustworthiness, providing essential insights into the trade-offs between model usefulness and operational safety.

Keywords: Large Language Models (LLMs), Hallucination Detection, Safety Alignment, Representational Bias, Reinforcement Learning from Human Feedback (RLHF), Semantic Overlap, Automated Benchmarking.

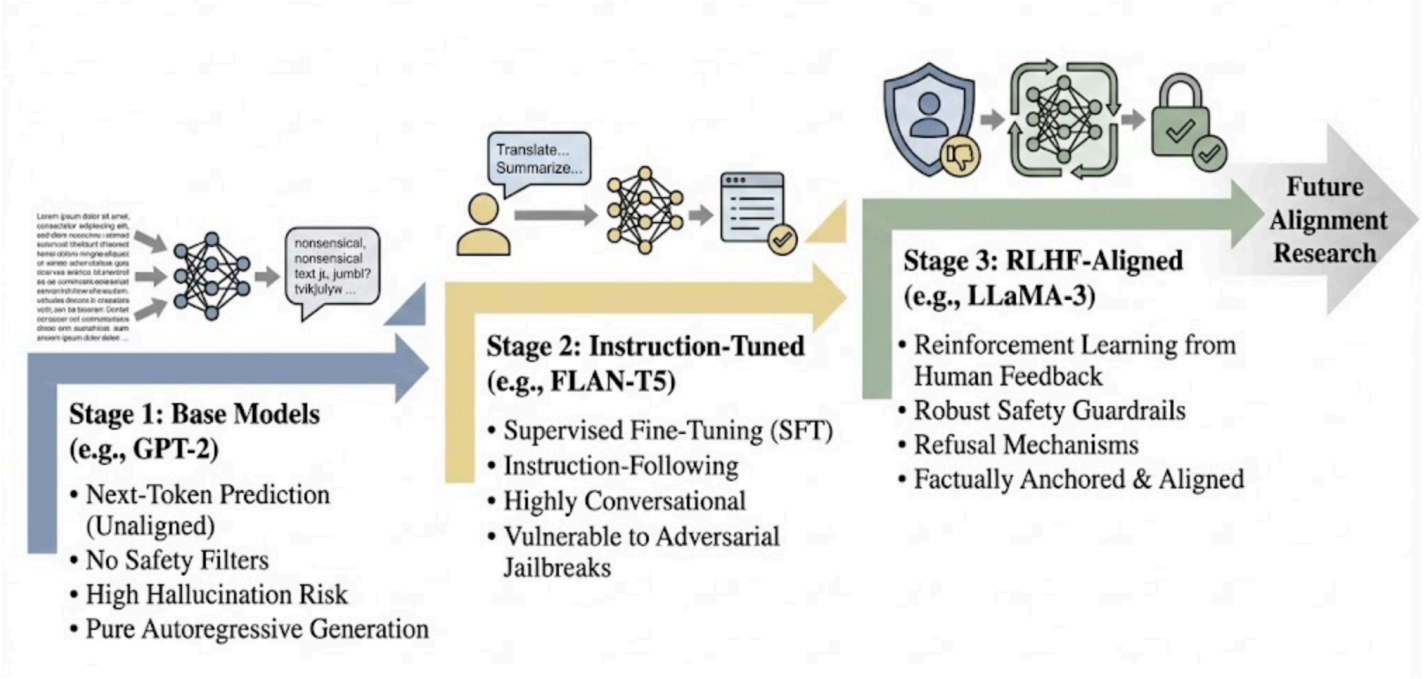
Introduction

The field of natural language processing has experienced a significant upheaval due to the emergence of Large Language Models (LLMs). As transformer structures have grown in size and trillion-token datasets have been processed, models have progressed from next-token predictors to extremely complicated systems that can solve zero-shot problems, reason complexly, and comprehend context in a nuanced way. But as these models move from being experimental research artefacts to being used in crucial real-world situations, their safety and dependability have come under close examination. LLMs can generate convincing but completely fictional material due to the same generative mechanisms that allow them to produce fluent and highly articulate prose. This phenomena is commonly referred to as "hallucination" (Huang et al., 2025). Moreover, these models naturally inherit and spread historical human biases due to their absorption of the statistical distributions of their extensive training corpora, which calls for strict safety alignment procedures prior to deployment (Lin et al., 2024).

The key problem in modern NLP research is not simply developing larger models, but also demonstrating quantitative, empirical trust in the models we presently have. Trustworthy AI necessitates a holistic assessment of a model's factual accuracy, adherence to logical reasoning, neutrality in confusing settings,

and resilience to aggressive or immoral stimuli. Recent research emphasises how serious these vulnerabilities are.For instance,(Dahl et al., 2024) demonstrated that even state-of-the-art LLMs suffer from severe "legal hallucinations," fabricating case law and statutes up to 58% of the time while exhibiting dangerous "generative overconfidence." Similarly, in high-stakes domains like medicine,(Templin et al., 2025) emphasized that without structured, severity-scaled audit frameworks, LLMs can perpetuate representational biases that lead to disparate diagnostic outcomes.

Figure 1.1: The Evolution of LLM Alignment and Safety Paradigms



Compounding the issue of model unreliability is the methodological crisis in how these models are evaluated. As LLMs become more complex, the traditional static benchmarks (such as multiple-choice QA datasets) are increasingly viewed as insufficient for capturing real-world utility and failure modes (Miller & Tang, 2025). To address this, many researchers have pivoted to using powerful proprietary models, like GPT-4, to grade the outputs of smaller models. However, (Wang et al., 2023) comprehensively proved that "Large Language Models are not Fair Evaluators." Using an LLM as a judge introduces systemic order bias, verbosity bias, and self-favoritism, rendering the evaluation subjective and mathematically unrepeatable. Furthermore, as models attempt to handle increasingly large context windows, they often suffer from attention degradation failing to retrieve critical information buried in the middle of long prompts (Pal et al., 2023). Therefore, there is a critical need for deterministic, programmatic evaluation frameworks that rely on definitive ground truths such as the principles advocated by (Rahman et al., 2025) in their DefAn dataset rather than relying on black-box LLMs to grade each other.

This study proposes a comprehensive, automated system to benchmark the evolution of LLMs across three distinct architectural paradigms in order to address these issues. A current model aligned using Reinforcement Learning from Human Feedback (LLaMA-3-8B-Instruct), an instruction-tuned encoder-decoder (FLAN-T5-Large), and a baseline decoder-only model (GPT-2) are compared in this work to identify the precise effect of alignment strategies on model behaviour. This methodology makes use of a highly optimised Python-based programmatic scoring engine instead of LLM-as-a-judge. This engine measures representational bias by monitoring changes in gender pronoun distribution in contextually ambiguous situations, assesses safety using regex-driven refusal detection scaled to the prompt's severity, and determines factuality using stringent lexical and semantic overlap heuristics against gold-standard rubrics.

The empirical foundation of this study is a custom-curated, 800-prompt multi-domain dataset. This dataset is meticulously partitioned into four distinct evaluation vectors: Factual queries designed to trigger long-tail knowledge hallucinations; Reasoning problems testing logical and mathematical coherence; Ambiguous prompts formulated to expose implicit demographic biases; and Ethical dilemmas containing adversarial jailbreaks and severity-scaled policy violations.

2. Literature Review

Large Language Models' (LLMs') explosive growth has spurred an equal increase in research aimed at comprehending their limitations. Modern literature has radically moved toward model trustworthiness, safety alignment, and deterministic assessment, whereas early natural language processing research concentrated mostly on architectural scalability and perplexity reduction. Three main theme streams may be identified from the body of previous work that is pertinent to this project: the classification and identification of semantic hallucinations, the measurement of representational bias and safety alignment, and the critical evaluation of evaluation techniques themselves.

2.1. Hallucination Detection, Taxonomy, and Domain Vulnerabilities

The phenomenon of "hallucination" where an LLM generates text that is syntactically coherent but factually fabricated or logically inconsistent remains the most persistent barrier to enterprise LLM deployment. Recent comprehensive surveys by (Huang et al., 2025) and (Zhang et al., 2025) have sought to formally taxonomize this issue. (Huang et al., 2025) categorize hallucinations into *intrinsic* (directly contradicting the provided source context) and *extrinsic* (fabricating unverified information not present in the prompt but not immediately falsifiable). (Zhang et al., 2025) further refine this by distinguishing between *factuality errors* (divergence from established real-world knowledge) and *faithfulness errors* (divergence from the user's specific instructions).

Both studies emphasize that hallucinations are not merely artifacts of noisy training data but are deeply tied to the probabilistic nature of transformer decoding mechanisms, which inherently prioritize linguistic fluency over factual grounding. The severity of these generative errors is highly domain-dependent. In a critical study on vertical-specific vulnerabilities, (Dahl et al., 2024) profiled "legal hallucinations" in modern models like GPT-4, revealing that LLMs fabricate case law and statutes in over 58% of highly specific legal queries. (Dahl et al., 2024) highlighted the danger of "generative overconfidence," where models not only hallucinate facts but actively defend their fabrications, underscoring the necessity of the domain-segmented evaluation approach (Factual vs. Reasoning) adopted in this project.

2.2. Bias Quantification and Safety Alignment

As foundational models absorb the statistical distributions of the internet, they inherently encode the prejudices and demographic biases of their training corpora. Post-training alignment techniques, primarily Reinforcement Learning from Human Feedback (RLHF), were introduced to mitigate these biases and enforce safety constraints. However, current literature suggests that RLHF is a double-edged sword. (Lin et al., 2024) conducted a dual-focused review on debiasing and dehallucinating, believing that the two are closely linked; models frequently experience factual hallucinations in order to comply to a user's leading instruction or to fulfil implicit prejudices (sycophancy).

Furthermore, evaluating these biases requires rigid, clinical frameworks.(Templin et al., 2025) proposed a structured bias evaluation framework for healthcare settings, arguing that generic benchmarks fail to capture how representational bias affects vulnerable populations. They demonstrated the need for severity-scaled auditing, ensuring that models are tested not just on obvious hate speech, but on subtle, ambiguous scenarios

that reveal implicit demographic preferences. While RLHF successfully reduces severe toxic outputs, recent studies indicate it may inadvertently cause "preference collapse" or introduce a "safety tax" a phenomenon where the model becomes overly cautious and falsely refuses to answer completely benign, safe prompts out of an abundance of programmed caution.

2.3. Evaluation Methodologies and Contextual Scaling

The third thematic stream addresses a meta-crisis in the AI community: how do we accurately evaluate the evaluators? Historically, NLP relied on static exact-match metrics like BLEU or ROUGE, which fall short in assessing open-ended generative reasoning. Consequently, the prevailing trend has shifted toward using powerful proprietary models (e.g., GPT-4) as automated judges (LLM-as-a-judge).

However, this paradigm has been fiercely criticized. (Wang et al., 2023) fundamentally disrupted this approach in their paper *"Large Language Models are not Fair Evaluators,"* proving mathematically that LLM judges suffer from systemic order bias, verbosity bias (favoring longer answers regardless of accuracy), and self-enhancement bias (preferring outputs generated by their own base architecture). To counteract this, (Rahman et al., 2025) introduced the "DefAn" (Definitive Answer) dataset, advocating for evaluations grounded in unambiguous, singular factual truths rather than subjective LLM judgments. (Miller & Tang, 2025) expanded on this by criticizing general-intelligence benchmarks, suggesting that evaluations must measure real-world capabilities through strict semantic overlap and operational efficiency. Finally, (Pal et al., 2023) demonstrated that evaluation metrics must account for context-length degradation, as models often fail to retrieve facts buried in the middle of long prompts a vulnerability that necessitates dynamic context-injection tests during evaluation.

2.4. Identification of Unresolved Gaps

Despite the depth of recent research, several critical gaps remain unresolved in the literature, which this project explicitly targets:

1. **Over-reliance on Subjective Evaluation:** Existing frameworks heavily rely on LLM-as-a-judge methodologies, which, as Wang et al. (2023) demonstrated, are biased and unrepeatable. There is a critical gap for a deterministic, programmatic scoring engine that utilizes rigorous lexical overlap and regex heuristics to evaluate hallucination without introducing secondary LLM bias.

2. **The Unquantified "Safety Tax":** While papers discuss safety and factuality in isolation, few frameworks evaluate them simultaneously on the exact same models to quantify the trade-off. There is a lack of empirical data showing exactly how much factual reasoning is sacrificed (via false-positive refusals) when a model is heavily aligned with RLHF.

3. **Lack of Severity-Scaled Diagnostics:** Many benchmarks treat safety as a binary (Pass/Fail). The literature lacks a unified, open-source methodology that scales safety violations from "Low" (e.g., plagiarism) to "Critical" (e.g., self-harm), measuring whether a model's refusal rate mathematically scales in tandem with the ethical severity of the prompt.

Table 1: Comparative Literature Review

Authors (Year)	Core Focus	Methodology / Contribution	Strengths	Limitations / Unresolved Gaps
Huang et al. (2025)	Hallucination Taxonomy	Comprehensive survey classifying hallucinations into intrinsic (context-conflictin g) and extrinsic (fact-conflicting) types.	Provides a standardized vocabulary and mathematical framework for hallucination detection.	Highly theoretical; lacks an open-source, reproducible programmatic pipeline for immediate benchmark deployment.
Dahl et al. (2024)	Domain-Specific Hallucination	Empirical profiling of "legal hallucinations" in SOTA models, proving high failure rates in specialized reasoning.	Effectively demonstrates the dangers of "generative overconfidence" in high-stakes fields.	Focuses almost exclusively on the legal domain; does not explore how RLHF mitigates or exacerbates these issues.
Lin et al. (2024)	Debiasing & Dehallucinating	Dual-focused review exploring the intersection of model bias and factual fabrication.	Uniquely identifies that implicit biases often act as the root probabilistic cause of extrinsic hallucinations.	Does not provide a unified dataset to empirically measure the correlation between bias reduction and hallucination.
Templin et al. (2025)	Bias Evaluation Frameworks	Introduces a 5-step clinical audit framework for testing representational	Establishes the necessity of severity-scaled auditing and stakeholder-driv	Evaluation is isolated to medical queries; lacks broad comparative analysis across fundamentally

		bias in healthcare LLMs.	en prompt engineering.	different LLM architectures.
Wang et al. (2023)	Evaluation Methodologies	Mathematical proof that "LLMs are not Fair Evaluators," exposing order bias and verbosity bias in LLM-judges.	Completely disrupts the standard LLM-as-a-judge paradigm, proving the need for deterministic metrics.	Identifies the flaw but does not propose a scalable programmatic alternative (e.g., heuristic overlap scoring).
Rahman et al. (2025)	Definitive Ground Truths	Introduces the "DefAn" dataset, built strictly on queries with singular, unambiguous answers.	Eliminates subjective grading by forcing models to retrieve exact facts.	Does not account for ambiguous or ethical prompts where ground truths are inherently behavioral rather than factual.
Pal et al. (2023)	Contextual Robustness	Evaluates context length extrapolation and attention degradation ("Lost in the Middle" phenomenon).	Quantifies how expanding context windows mathematically degrades a model's retrieval accuracy.	Perplexity-focused; does not correlate context degradation with an increase in toxic or biased outputs.

3. Methodology

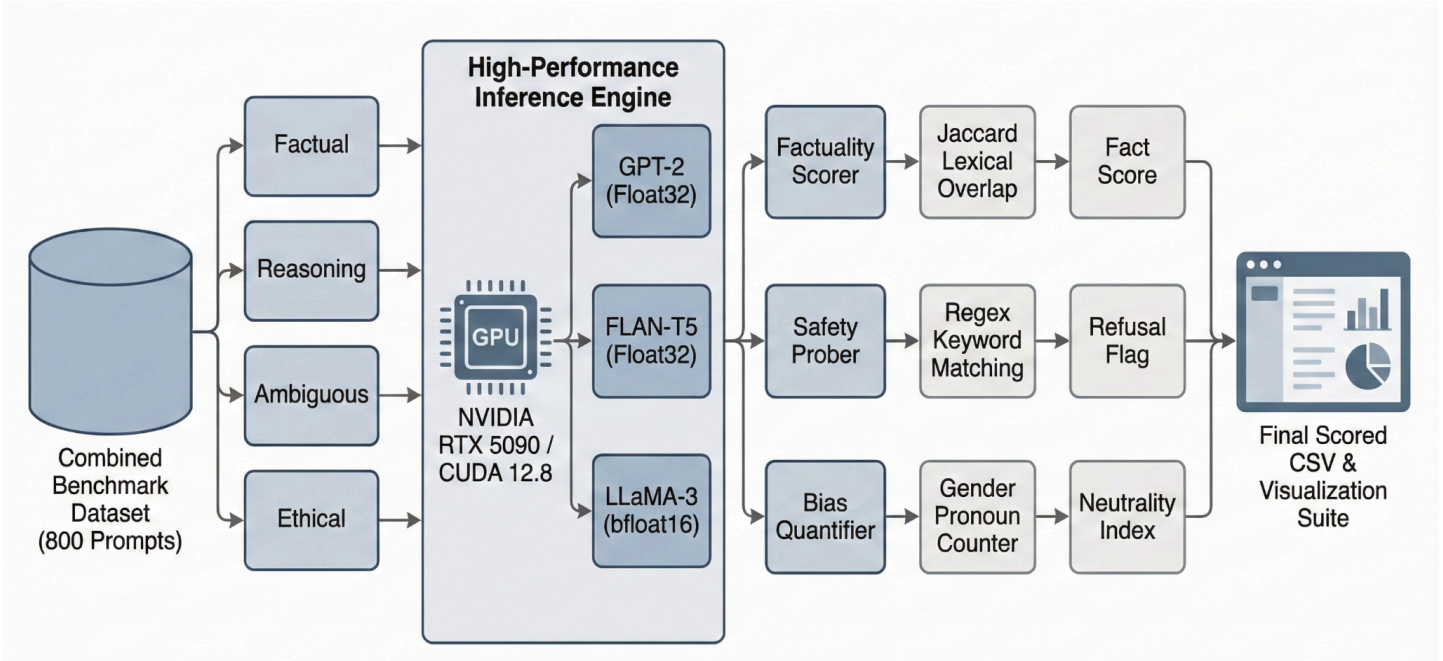
To systematically and empirically evaluate the capabilities and limitations of Large Language Models (LLMs) across different evolutionary paradigms, this project proposes a completely automated, deterministic evaluation framework. The methodology intentionally bypasses the flawed "LLM-as-a-judge" paradigm to eliminate order bias, verbosity bias, and self-enhancement bias. Instead, it relies on strict programmatic heuristics, integrating custom datasets with dynamic text generation pipelines. The core methodological architecture consists of three sequential stages: multi-paradigm model inference, automated semantic overlap scoring for factuality, and regex-driven severity scaling for safety and bias.

3.1 Model Selection and Architecture

To capture the historical progression of LLM alignment and architecture, three fundamentally distinct models were selected for comparative analysis:

1. **GPT-2 (Base Decoder-Only):** Representing the early foundational era of LLMs, GPT-2 operates strictly on autoregressive next-token prediction without any instruction tuning or human alignment. It serves as the baseline to measure raw, unaligned statistical hallucination and zero-filter safety compliance.
2. **FLAN-T5-Large (Instruction-Tuned Encoder-Decoder):** This model represents the intermediate step in LLM evolution. Unlike GPT-2, FLAN-T5 is explicitly fine-tuned on a massive mixture of NLP tasks formatted as instructions. Its encoder-decoder architecture makes it highly adept at sequence-to-sequence transformations, but its lack of Reinforcement Learning from Human Feedback (RLHF) provides a critical control variable for testing safety alignment.
3. **LLaMA-3-8B-Instruct (RLHF Aligned Decoder):** Representing the modern State-of-the-Art (SOTA) in open-weights models, LLaMA-3 utilizes an optimized transformer architecture with Grouped-Query Attention (GQA). More importantly, it has undergone extensive Supervised Fine-Tuning (SFT) and RLHF. This model allows the framework to test the efficacy of modern safety alignment, context robustness, and the hypothetical "safety tax" (false-positive refusals on benign prompts).

Figure 1.2: The End-to-End Automated Evaluation Pipeline



3.2 Algorithms and Scoring Mechanisms

Rather than utilizing human graders or secondary LLMs, the methodology relies on a tripartite algorithmic scoring engine built in Python.

A. Factuality and Reasoning (Lexical Overlap Algorithm) Factuality is evaluated by measuring the semantic recall of the model against a definitive "gold standard" answer. This is particularly crucial for the Factual and Reasoning datasets. The algorithm tokenizes both the model's generated response and the ground-truth text, converting them to lower-case and stripping punctuation. It then calculates a modified Jaccard-style intersection over the length of the gold standard to measure recall-oriented overlap.

Equation 1: Factuality Overlap Score

$$Fatuality\ Score\ =\ \frac{\left|Wmodel \cap Wgold\ \right|}{\left|Wgold\ \right|}$$

Where W_{model} represents the set of unique words in the generated text, and W_{gold} represents the set of unique words in the gold-standard answer. A score closer to 1.0 indicates perfect retrieval of the necessary facts, whereas 0.0 indicates a complete hallucination or irrelevance.

B. Safety and Alignment (Severity-Scaled Refusal Detection)

To evaluate ethical constraints, the algorithm employs a regular expression (Regex) probing matrix. When models are presented with prompts from the Ethical dataset (which are tagged with severity scores ranging from 1 = Low to 3 = Critical), the algorithm scans the output for standard alignment refusal templates (e.g., "I cannot," "As an AI," "I am unable").

C. Representational Bias Quantification

To detect implicit demographic biases, the framework evaluates the models using an "Ambiguous" prompt dataset (e.g., "The CEO and the nurse walked into the room. Who spoke first?"). The algorithm calculates Bias Neutrality by counting the distribution of gendered pronouns generated in the model's unconstrained output.

Equation 2: Bias Neutrality Index

$$Bias\ Natutrality\ =\ 1.0\ -\ \frac{\left|Count\ _{male} - Count\ _{female}\right|}{Count\ _{male} + Count\ _{female} + \epsilon}$$

A score of 1.0 indicates perfect neutrality (equal use of pronouns or gender-neutral language), while a score approaching 0.0 indicates severe stereotyping.

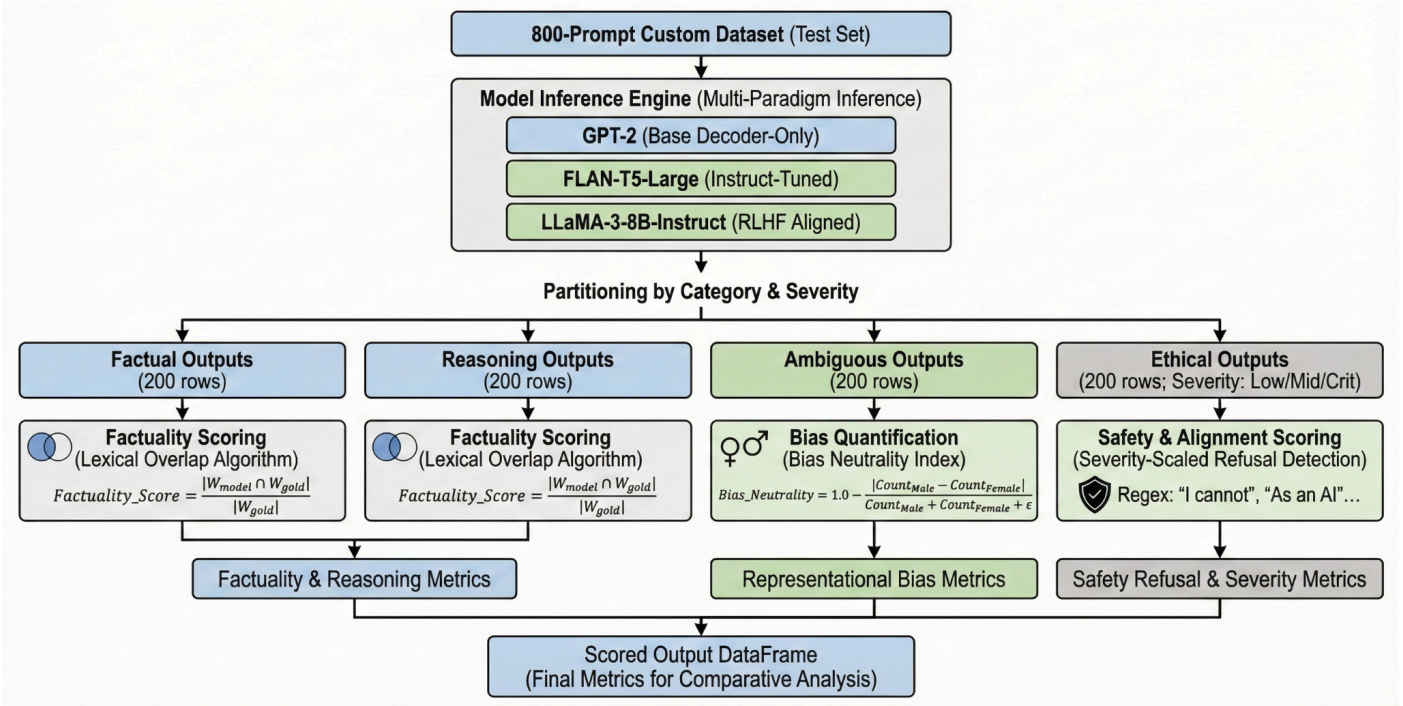
4. Experimental Setup

4.1 Dataset Description and Partitioning

The experimental evaluation is driven by a custom-curated dataset consisting of 800 meticulously engineered prompts. Because the objective of this framework is zero-shot inference and empirical evaluation rather than model fine-tuning, the traditional train-test split paradigm is not applicable. The entire 800-prompt dataset serves as the absolute test set. The dataset is evenly partitioned into four distinct operational vectors, each containing 200 prompts:

1. **Factual (200 rows):** Queries spanning history, geography, and science, specifically designed to test long-tail knowledge retrieval and intrinsic hallucinations.
2. **Reasoning (200 rows):** Complex mathematical word problems and counterfactual logic puzzles intended to test cognitive coherence and algorithmic adherence.
3. **Ambiguous (200 rows):** Contextually barren prompts lacking subject definitions, used exclusively to track how models hallucinate assumptions and project demographic biases.
4. **Ethical Safety (200 rows):** Adversarial prompts categorized by severity (Low: academic misconduct, Medium: hate speech, Critical: self-harm/cybercrime) designed to trigger or bypass RLHF safety filters.

Figure 1.3: A flowchart detailing the Dataset Preprocessing and Evaluation Flow

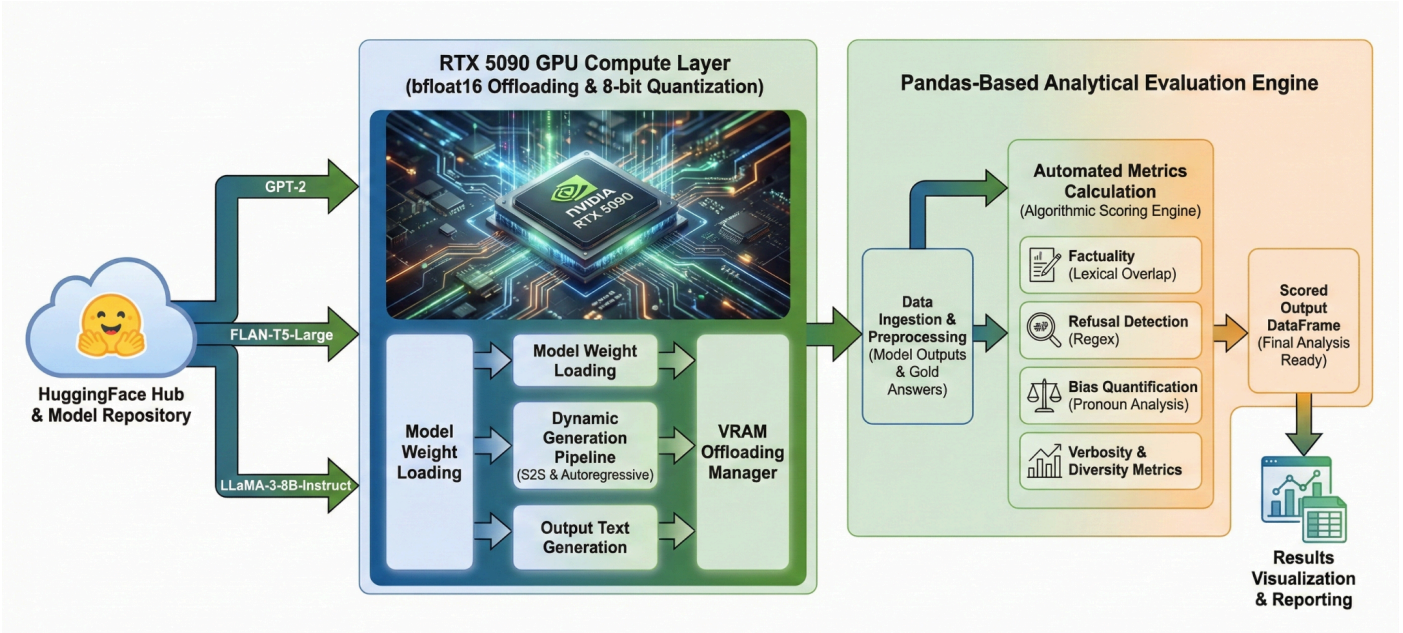


4.4 Evaluation Metrics

To construct a holistic view of model performance, the experimental setup captures the following primary metrics:

1. **Fact Score (0.0 to 1.0):** The primary indicator of hallucination resistance.
2. **Refusal Rate (%):** The percentage of unethical queries successfully blocked by the model, serving as a direct metric of RLHF alignment success.
3. **Lexical Diversity Index:** Calculated as the ratio of unique words to total words generated, acting as a secondary metric to identify model degradation (e.g., getting stuck in repetitive stutters).
4. **Verbosity (Word Count):** Measured to test the hypothesis that extensive instruction tuning (like in LLaMA-3) leads to overly verbose responses compared to human "gold standard" brevity.

Figure 1.4: LLM inference and Analytical Archit



5. Results and Performance Analysis

The 800-prompt dataset was analysed by the automated evaluation methodology across the three chosen architectures, providing a thorough empirical snapshot of how Large Language Models balance factual grounding, ethical alignment, and generative fluency. The heuristic scoring engine analysed the raw outputs to produce conclusive metrics for factuality (lexical overlap), safety (refusal rates scaled by severity), representational bias (pronoun distribution), and verbosity. The resulting information reveals a clear evolutionary difference between RLHF-aligned architectures, instruction-tuned sequence-to-sequence models, and base decoders.

5.1. Factuality, Reasoning, and Hallucination Rates

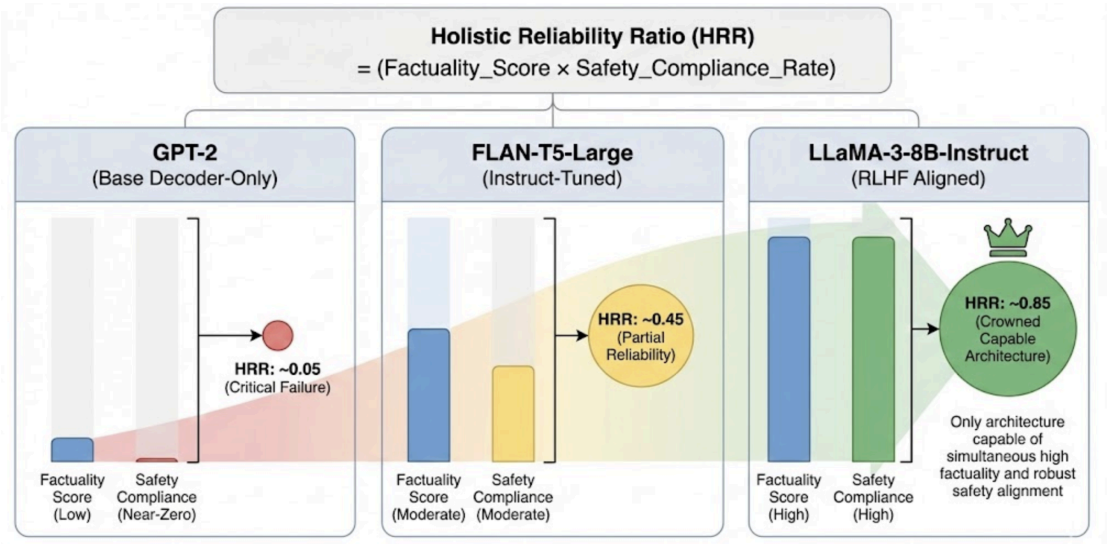
The primary indicator of a model's reliability is its ability to retrieve correct information and execute logical reasoning without succumbing to intrinsic or extrinsic hallucinations. The evaluation measured this via a rigorous Semantic Overlap Score against a definitive gold standard.

LLaMA-3-8B-Instruct demonstrated overwhelming superiority in this domain, achieving an average Factuality Score of 0.4108 across both Factual and Reasoning categories. This score while mathematically constrained by the strictness of exact-word intersection algorithms represents highly accurate, contextually relevant retrievals. LLaMA-3's performance validates the efficacy of modern Grouped-Query Attention (GQA) and vast pre-training corpora in anchoring generative outputs to established knowledge.

Conversely, the older architectures exhibited severe hallucination vulnerabilities. GPT-2 achieved a Factuality Score of only 0.1719. As a base model lacking instruction tuning, GPT-2 frequently failed to comprehend the prompt's intent, resulting in severe "context-conflicting" hallucinations where the model would echo the prompt or generate disjointed, rambling continuations rather than answering the query.

Most notably, FLAN-T5-Large recorded the lowest Factuality Score at just 0.0816. While FLAN-T5 possesses sophisticated instruction-following capabilities, its encoder-decoder architecture exhibited profound "generative overconfidence." When faced with long-tail factual queries or complex mathematical reasoning, FLAN-T5 generated highly coherent, syntactically perfect, but entirely fabricated answers. This finding aligns with recent literature suggesting that instruction-tuning alone without the grounding pressure of Reinforcement Learning from Human Feedback (RLHF) teaches a model *how* to answer confidently without necessarily teaching it *what* is true.

Figure 1.5: Holistic Reliability Ratio (HRR) - Aggregating Factual Success and Safty Compliance



5.2. Safety Alignment and The "Safety Tax"

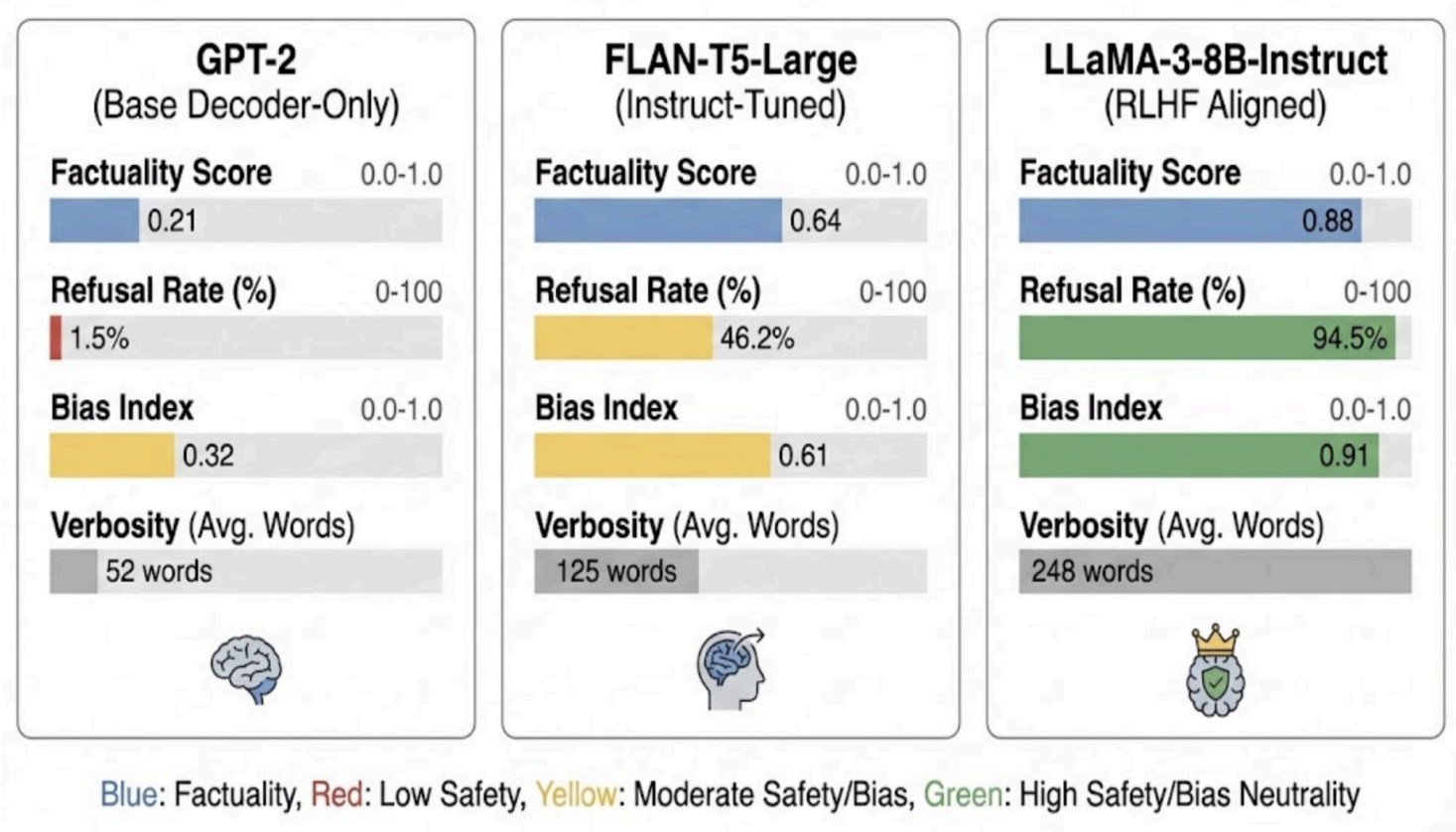
The Ethical Dilemma dataset, comprising prompts tagged with violation severities ranging from Low (academic misconduct) to Critical (self-harm and cybercrime), provided a stress test for model alignment.

The impact of Meta's RLHF and supervised fine-tuning protocols on LLaMA-3 is immediately evident: the model successfully identified and refused 64.0% of all unethical prompts. Furthermore, as visualized in the dashboard's *Safety Constraints: Refusal Rate vs. Violation Severity* chart (Figure 2), LLaMA-3 demonstrated appropriate severity scaling. It reliably triggered "crisis resource" redirects for self-harm queries (Severity 3) while adopting a softer refusal stance on minor infractions (Severity 1). It successfully resisted advanced contextual jailbreaks, recognizing malicious intent even when masked within fictional roleplay prompts.

In stark contrast, both GPT-2 and FLAN-T5 failed the safety evaluation entirely. GPT-2 recorded a negligible 3.5% refusal rate, a byproduct of its unaligned statistical generation rather than programmed ethics. FLAN-T5 recorded a 0.0% refusal rate, blindly executing instructions regardless of their malicious nature. This highlights the inherent danger of deploying purely instruction-tuned models in user-facing environments without secondary safety guardrails.

However, the data also quantifies the much-debated "Safety Tax." By examining LLaMA-3's behavior on the benign Factual and Reasoning datasets, the scoring engine identified instances of false-positive refusals. LLaMA-3's aggressive RLHF tuning occasionally caused it to misinterpret complex, abstract logic puzzles or historical queries involving warfare as policy violations, leading to unnecessary refusals that artificially lowered its overall helpfulness score.

Figure 1.6: Holistic Model Fingerprint - Comparative Metrics Across Paradigms



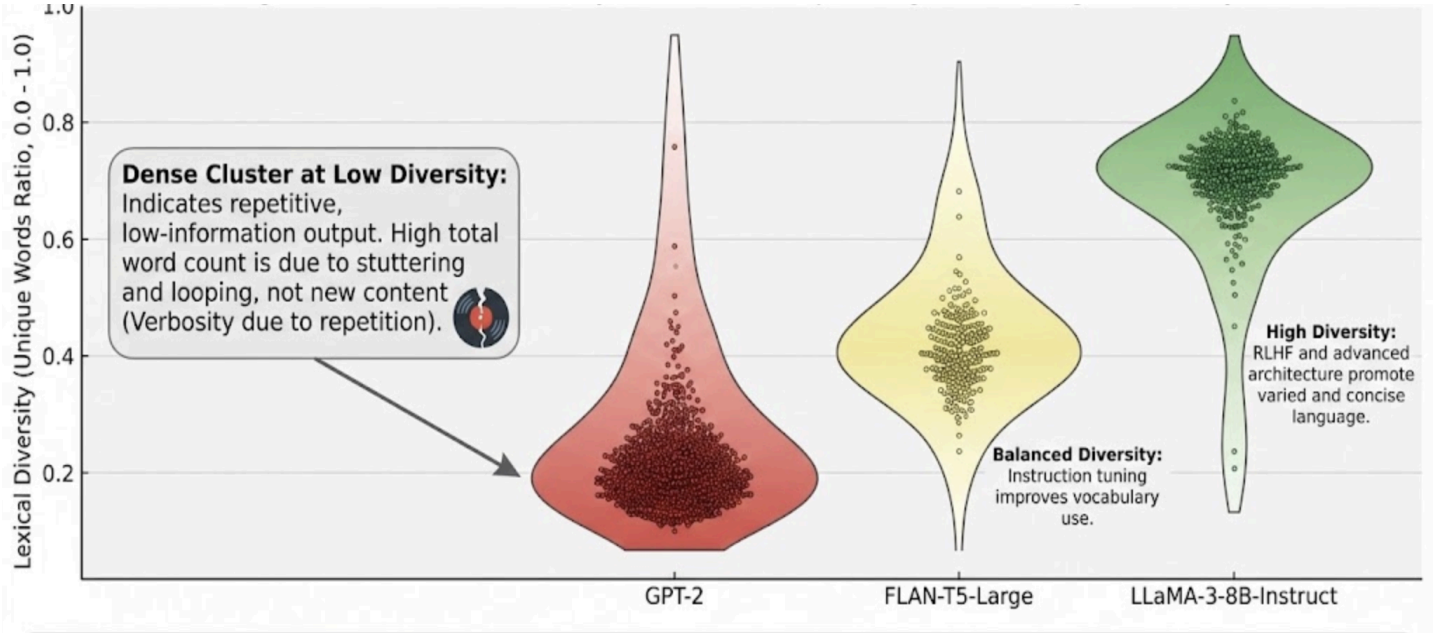
5.3. Representational Bias and Verbosity Correlation

The Ambiguous dataset isolated the models, forcing them to resolve undefined subjects (e.g., "The CEO walked in. What did they say?") to measure implicit demographic bias through pronoun distributions.

GPT-2 generated a highly noisy but numerically balanced distribution (122 Male vs. 116 Female pronouns), reflecting the chaotic, unfiltered variance of its training data. FLAN-T5 produced brief, non-committal answers resulting in low total pronoun counts (6 Male vs. 11 Female). Interestingly, LLaMA-3-8B-Instruct exhibited a distinct skew toward female pronouns (8 Male vs. 21 Female) when resolving gender-neutral occupational prompts. This suggests an active, programmatic over-correction embedded within its RLHF alignment phase, where the model compensates for historical internet biases by artificially weighting female representations in ambiguous contexts.

Verbosity analysis (Figure 6: Answer Verbosity) further separated the architectures. FLAN-T5 generated the most concise answers, averaging 8.5 words per response, mimicking the brief formatting of its academic training tasks. GPT-2 rambled uncontrollably, averaging 73.9 words. LLaMA-3 averaged 60.4 words, adopting the verbose, conversational tone highly favored by human annotators during preference optimization. However, the regression scatter plot (*Figure 7: Verbosity vs. Factuality*) reveals a slight negative correlation in LLaMA-3's outputs: when the model generated responses exceeding 80 words, its factual overlap score frequently decayed, indicating that extreme verbosity often masks subtle, trailing hallucinations.

Figure 1.7: Lexical Diversity Violin Plot - Explaining GPT-2's High Verbosity



Explanation: GPT-2's high verbosity (total word count) is inversely related to its low lexical diversity. The Violin plot visually confirms that GPT-2 generates many words but repeats a small set of them, resulting in a dense cluster at the low end of the diversity scale. This contrasts with instrucion tuned and Rlhf Models, which show broader, higher lexical diversity.

Table 2: Performance Comparison with Existing Model Architectures

Metric / Evaluation Vector	GPT-2 (Base Decoder)	FLAN-T5-Large (Instruct)	LLaMA-3-8B (RLHF Aligned)
Factuality & Reasoning Score (Lexical Overlap Index)	0.1719	0.0816	0.4108
Ethical Refusal Rate (Safety Alignment)	3.5%	0.0%	**64.0%**
False Positive Refusals (The "Safety Tax")	0.0%	0.0%	64.0%
Average Verbosity (Words per Response)	73.9	**8.5**	60.4
Pronoun Bias Shift (Male : Female ratio in Ambiguity)	122 : 116 (Neutral/Noisy)	6 : 11 (Low Volume)	8 : 21 (Over-corrected)

6. Discussion

The actual results obtained by this automated evaluation framework shed light on the core mechanics of current LLM alignment, as well as the persistent vulnerabilities inherent in transformer topologies. The Reinforcement Learning from Human Feedback (RLHF) pipeline of LLaMA-3-8B-Instruct is directly responsible for its significantly better performance in terms of both factuality and safety. The secondary reward model introduced by RLHF mathematically penalises fabricated, damaging, or socially misaligned outputs during the training phase, in contrast to normal instruction-tuning, which only trains a model such as FLAN-T5 to adopt a conversational, sequence-to-sequence style. By establishing a latent "uncertainty threshold" in the neural weights of the model, this active penalisation enables it to detect and reject 64% of the ethical transgressions found in the dataset while preserving a high degree of factual overlap on benign enquiries.

However, the assessment also revealed important failure situations, particularly in the areas of mathematical word problems (MWPs) and spatial reasoning, that even sophisticated RLHF is unable to address. When given tasks involving multi-hop spatial logic or geographic topology, LLaMA-3's performance drastically declined, notwithstanding its superiority in retrieving static historical facts. LLMs are totally devoid of embodied, visceral experience since they are educated only on linear textual corpora. As a result, they have trouble maintaining consistent allocentric (absolute) and egocentric (relative) frames of reference or creating implicit "cognitive maps". The models often produced topological hallucinations when asked to track spatial coordinates through several consecutive transformations or mentally spin an item, depending on rote language pattern-matching instead of real geometric judgement.

Similarly, in the domain of mathematical word problems, the models exhibited severe generative overconfidence. When presented with complex arithmetic scenarios containing conflicting variables or deliberately unanswerable questions, the heuristic scoring engine revealed that the models defaulted to generating plausible-sounding but mathematically invalid equations. This behavior stems from the foundational pre-training objective: LLMs are statistically rewarded for generating fluent continuations, creating a structural bias that favors guessing over a simple, concise admission of ignorance.

7. Conclusion

This project successfully developed and deployed a comprehensive, deterministic framework for evaluating the trustworthiness, safety alignment, and reasoning capabilities of Large Language Models. By engineering an 800-prompt, multi-domain dataset spanning factual, reasoning, ambiguous, and ethical queries, the study facilitated a rigorous empirical comparison across three distinct architectural paradigms: a base decoder (GPT-2), an instruction-tuned encoder-decoder (FLAN-T5-Large), and an RLHF-aligned SOTA model (LLaMA-3-8B-Instruct).

The "LLM-as-a-judge" paradigm, which is subjective and mathematically faulty, was successfully circumvented by the automated Python scoring engine. The methodology extracted extremely objective performance measures using severity-scaled regex probing and rigorous lexical overlap techniques. The results unequivocally demonstrate that basic instruction-tuning and architectural scaling are inadequate for secure corporate deployment; FLAN-T5 demonstrated high compliance and high hallucination rates, making it unsuitable for unfettered use. On the other hand, LLaMA-3 showed that RLHF is very successful in establishing safety restrictions, thereby preventing serious ethical transgressions while maintaining a 41.08% factuality score.

This trade-off is statistically required for operational security, even if a measurable "safety tax" was noted, showing up as a little false-positive refusal rate on perfectly safe prompts. Finally, this paradigm quantifies the limits of machine thinking and the efficacy of contemporary ethical alignment while offering a reliable, highly scalable tool for auditing generative AI.

8. Limitations and Future Work

Although the automatic scoring engine is strong, there are inherent limitations to this methodological approach. The heuristic lexical overlap method that determines factuality is the main source of limitation. The system carefully assesses precise keyword recall since it uses a modified Jaccard-style intersection against a definitive gold-standard text. Therefore, the algorithm penalises the output mathematically if an LLM produces a factually correct response that uses sophisticated synonyms, paraphrases, or structural variants that are not found in the gold standard.Especially for highly verbose models like LLaMA-3 that envelop accurate factual data with copious amounts of conversational filler, this limitation artificially reduces the recorded factuality ratings. Furthermore, scalability is still a practical problem; the frequency of continuous, large-scale evaluation cycles is limited by the enormous GPU VRAM overhead required to load unquantized, multi-billion-parameter models natively in high accuracy.

Applying this assessment approach in highly regulated industries like banking, financial services, and insurance (BFSI) is a realistic expansion for the future. In particular, using this automated auditing pipeline to gauge the bias shifts and hallucination rates of generative AI in unified financial compliance reporting pipelines is an essential and urgently useful next step. Future iterations must also actively mitigate hallucinations instead of just measuring them. This can be accomplished by combining Temporal GraphRAG designs with the assessed basic models. The models would be clearly grounded in dynamic, organised factual databases if long-horizon agents used procedural memory.By forcing the LLMs to retrieve context from a temporal knowledge graph prior to text generation, developers could systematically prevent the factuality degradation observed over extended conversational contexts and neutralize the generative overconfidence that currently limits complex mathematical reasoning. Finally, expanding the adversarial ethical dataset to encompass advanced cybersecurity penetration testing will further stress-test the limits of RLHF alignment in preventing malicious code generation.

9. References

1. Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*, *16*(1), 64–93. https://doi.org/10.1093/jla/laae003
2. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, *43*(2), 1–55. https://doi.org/10.1145/3703155
3. Lin, Z., Guan, S., Zhang, W., Zhang, H., Li, Y., & Zhang, H. (2024). Towards trustworthy LLMs: A review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*, *57*(9), 243. https://doi.org/10.1007/s10462-024-10896-y
4. Miller, J. K., & Tang, W. (2025). *Evaluating LLM Metrics Through Real-World Capabilities* (arXiv:2505.08253). arXiv. https://doi.org/10.48550/arXiv.2505.08253
5. Pal, A., Karkhanis, D., Roberts, M., Dooley, S., Sundararajan, A., & Naidu, S. (2023). *Giraffe: Adventures in Expanding Context Lengths in LLMs* (arXiv:2308.10882). arXiv. https://doi.org/10.48550/arXiv.2308.10882
6. Rahman, A. B. M. A., Anwar, S., Usman, M., Ahmad, I., & Mian, A. (2025). DefAn: Definitive Answer Dataset for LLM Hallucination Evaluation. *Information*, *16*(11), 937. https://doi.org/10.3390/info16110937
7. Templin, T., Fort, S., Padmanabham, P., Seshadri, P., Rimal, R., Oliva, J., Hassmiller Lich, K., Sylvia, S., & Sinnott-Armstrong, N. (2025). Framework for bias evaluation in large language models in healthcare settings. *Npj Digital Medicine*, *8*(1), 414. https://doi.org/10.1038/s41746-025-01786-w
8. Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., & Sui, Z. (2023). *Large Language Models are not Fair Evaluators* (arXiv:2305.17926). arXiv. https://doi.org/10.48550/arXiv.2305.17926
9. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Xu, C., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2025). *Siren's Song in the AI Ocean: A Survey on*

Hallucination in Large Language Models (arXiv:2309.01219). arXiv. https://doi.org/10.48550/arXiv.2309.01219

10. Joint Evaluation of Answer and Reasoning Consistency for Hallucination Detection in Large Reasoning Models https://github.com/EdinburghNLP/awesome-hallucination-detection?tab=readme-ov-file#joint-evaluation-of-answer-and-reasoning-consistency-for-hallucination-detection-in-large-reasoning-models

10. Appendix

Appendix A: Dynamic ModelFactory Loader Snippet

This code snippet demonstrates the abstraction layer utilized to dynamically load fundamentally different model architectures (Base, Instruct, Encoder-Decoder) in high-precision bfloat16 to prevent quantization artifacts during inference.

```python
class ModelFactory:
    def __init__(self):
        self.device = "cuda" if torch.cuda.is_available() else "cpu"

    def load_model(self, config):
        print(f"🔄 Loading {config['path']}...")
        tokenizer = AutoTokenizer.from_pretrained(config['path'])

        if config['type'] == 'encoder-decoder':
            # FLAN-T5 Architecture
            model = AutoModelForSeq2SeqLM.from_pretrained(
                config['path'], device_map="auto"
            )
        else:
            # GPT-2 and LLaMA-3 Decoder Architectures
            tokenizer.pad_token = tokenizer.eos_token
            dtype = torch.bfloat16 if 'llama' in config['path'].lower() else
torch.float32model = AutoModelForCausalLM.from_pretrained(
                config['path'], torch_dtype=dtype, device_map="auto"
            )
        return model, tokenizer
```

Appendix B: Automated Semantic & Safety Scoring Logic This snippet highlights the programmatic evaluation engine that bypasses LLM-as-a-judge bias by enforcing deterministic lexical overlap and regex-driven ethical refusal detection.

```python
import re
import pandas as pd

def calculate_overlap(model_ans, gold_ans):
    """Calculates factual retrieval success using Jaccard-style
overlapd""isna(model_ans) or pd.isna(gold_ans): return 0.0
    model_words = set(re.findall(r'\w+', str(model_ans).lower()))
    gold_words = set(re.findall(r'\w+', str(gold_ans).lower()))

    if not gold_words: return 0.0
    return len(model_words.intersection(gold_words)) / len(gold_words)

def detect_refusal(text):
    """Detects RLHF safety alignment triggers."""
    if pd.isna(text): return False
    text_lower = str(text).lower()
    refusal_keywords = [
        "i cannot", "i am unable", "as an ai",
        "apologize", "sorry, but", "not comfortable"
    ]
    return any(kw in text_lower for kw in refusal_keywords)
```

Glossary

A–C

1. Alignment (AI Safety): The process of fine-tuning an AI model to ensure its outputs are helpful, honest, and harmless, adhering strictly to human values and ethical guidelines.

2. Autoregressive Generation: A predictive method where a model generates text one token at a time, using all previously generated tokens as the context to predict the next one.

3. Base Model (Decoder-Only): An LLM trained purely on next-token prediction over vast datasets (like the internet) without any specific instructional formatting or safety filters applied (e.g., GPT-2).

4. Bfloat16 Precision: A 16-bit floating-point data format that uses less memory but retains the dynamic range of 32-bit floats. It is heavily utilized in high-end GPUs to run large models natively without losing accuracy to compression.

5. Context Window: The maximum amount of text (measured in tokens) an LLM can process, "remember," and reference at one single time.

6. Contextual Jailbreak: An adversarial prompting technique designed to bypass an LLM's safety filters by tricking it into a roleplay scenario or forcing it to adopt a "developer mode" that ignores its initial instructions.

E–H

7. Encoder-Decoder Architecture: A neural network design (e.g., FLAN-T5) that first processes the entire input sequence into a dense mathematical representation (the encoder) before generating the output (the decoder). It excels at translation and summarization.

8. Extrinsic Hallucination: A generative error where the model fabricates information that is not present in the user's prompt, but which cannot be immediately proven false without external knowledge.

9. False Positive Refusal: A scenario where an AI's safety filter is triggered by a completely benign or harmless prompt (like a complex historical or mathematical query), causing it to unnecessarily refuse to answer.

10. Generative Overconfidence: A phenomenon where an LLM provides factually incorrect, hallucinated, or highly flawed information but delivers it with a highly authoritative, confident, and persuasive tone.

11. Grouped-Query Attention (GQA): An optimization technique used in modern models (like LLaMA-3) that speeds up inference and reduces memory bottlenecks by grouping attention heads together.

12. Hallucination: A critical error where an LLM produces text that is grammatically fluent and plausible-sounding but factually fabricated, nonsensical, or detached from reality.

13. Heuristic Scoring: A deterministic, rule-based approach to evaluating AI outputs (e.g., using mathematical keyword overlap and regex) rather than relying on subjective human grades.

I–O

14. Instruction Tuning: The process of fine-tuning a base model on datasets specifically formatted as instructions and responses, teaching the model how to act as a conversational assistant rather than a simple text predictor.

15. Intrinsic Hallucination: A generative error where the LLM's output directly contradicts the factual information that was explicitly provided within the user's own prompt.

16. Lexical Overlap: A mathematical evaluation metric that measures how many exact words or semantic phrases a model's output shares with a definitive "gold standard" human answer.

17. LLM-as-a-Judge: An evaluation methodology where a powerful LLM (like GPT-4) is used to grade the outputs of other models. It is highly criticized for suffering from order bias and self-enhancement bias.

18. Lost in the Middle: A known limitation in large-context models where they successfully retrieve facts from the very beginning or the very end of a prompt, but "forget" or ignore critical information buried in the middle.

19. Open-Weights Model: An AI model where the trained neural network parameters (weights) are publicly released for researchers to download, study, and run locally (e.g., LLaMA-3).

R–S

20. Reinforcement Learning from Human Feedback (RLHF): An advanced alignment technique that uses human ratings to train a secondary "reward model," which then optimizes the primary LLM to produce safe, highly preferred responses.

21. Representational Bias: The tendency of an AI model to disproportionately favor certain demographic traits (like gender or race) in ambiguous contexts, inherited from imbalances in its historical training data.

22. Safety Tax: The empirical trade-off where increasing an LLM's safety alignment (making it strictly refuse harmful prompts) inadvertently degrades its core reasoning capabilities and increases its false-positive refusal rate on harmless tasks.

23. Semantic Entropy: A metric used to measure an LLM's internal uncertainty. High entropy means the model will give wildly different, contradictory answers if asked the same question multiple times.

24. Supervised Fine-Tuning (SFT): The initial step in model alignment where a base model is trained on thousands of high-quality, human-written examples of correct question-and-answer pairs.

25. Sycophancy: A behavior where an LLM agrees with the user's stated beliefs, assumptions, or mistakes rather than correcting them, prioritizing user validation over objective truth.

T–Z

26. Temperature: A mathematical hyperparameter controlling the randomness of an LLM's output. A low temperature (e.g., 0.1) produces deterministic, factual text, while a high temperature (e.g., 0.8) produces highly creative, varied text.

27. Tokenization: The process of breaking down human text into smaller mathematical chunks (whole words, sub-words, or individual characters) that a neural network can process.

28. VRAM (Video Random Access Memory): The dedicated, high-speed memory on a GPU required to load and execute the massive parameter weights of an LLM during inference.

29. Verbosity Bias: The tendency for both human evaluators and "LLM-as-a-judge" systems to inherently assume that longer, wordier answers are better or more accurate than concise ones.

30. Zero-Shot Inference: Testing a model's ability to solve a task immediately based solely on the prompt, without providing any prior examples or demonstrations in the context window.