Enhancing Large Language Models: A Survey on Process Supervision and AI Alignment

Abstract

Large Language Models (LLMs) are pivotal in advancing artificial intelligence, particularly in natural language processing (NLP). This survey explores the enhancement of LLMs through process supervision and AI alignment, emphasizing their significance in aligning AI systems with human values. Process supervision techniques, including step-level supervision and Process Reward Models (PRMs), optimize LLM performance by providing structured feedback and improving reasoning capabilities. AI alignment ensures LLMs operate ethically, addressing challenges such as bias and ethical compliance. The integration of reinforcement learning further refines LLM outputs, aligning them with human preferences. Despite advancements, challenges persist, including the integration of human preferences and the reliability of LLM-generated annotations. The survey highlights the importance of verification methods, such as step-by-step and process verifiers, in ensuring LLM accuracy and reliability. Future research should focus on refining model architectures, enhancing process supervision techniques, and expanding LLM applications across diverse domains. By addressing these areas, LLMs can achieve technical excellence while adhering to ethical and societal standards, fostering trust and reliability in AI systems.

1 Introduction

1.1 Significance of LLMs in AI

Large Language Models (LLMs) are integral to the advancement of artificial intelligence, significantly enhancing natural language processing (NLP) and enabling applications such as keyword extraction [1] and public affairs analysis, which foster transparency and accountability [2]. Their adeptness at interpreting ambiguous user intents allows for the generation of relevant visualizations, thus improving user experience [3].

LLMs are increasingly integrated with automatic speech recognition (ASR) systems, improving performance in tasks like Mandarin ASR [4]. They also address complex challenges in the digital space, such as spam email detection, where traditional methods often struggle against evolving phishing tactics [5]. Furthermore, LLMs show promise in enhancing embodied AI tasks, although current approaches face limitations [6]. Research employing cognitive science techniques has revealed latent capabilities of LLMs, enriching our understanding of their potential [7]. A systematic review of tool learning with LLMs highlights the need to bridge knowledge gaps regarding their advantages and implementation [8].

Despite these advancements, challenges remain, particularly in integrating human preferences into LLMs to align AI systems with human values [9]. The rapid generation of annotations by LLMs also raises concerns about reliability, emphasizing their critical role in AI and data annotation processes [10]. As research continues to optimize LLM performance and expand their applications, their significance in addressing complex challenges and fostering innovation across sectors becomes increasingly evident.

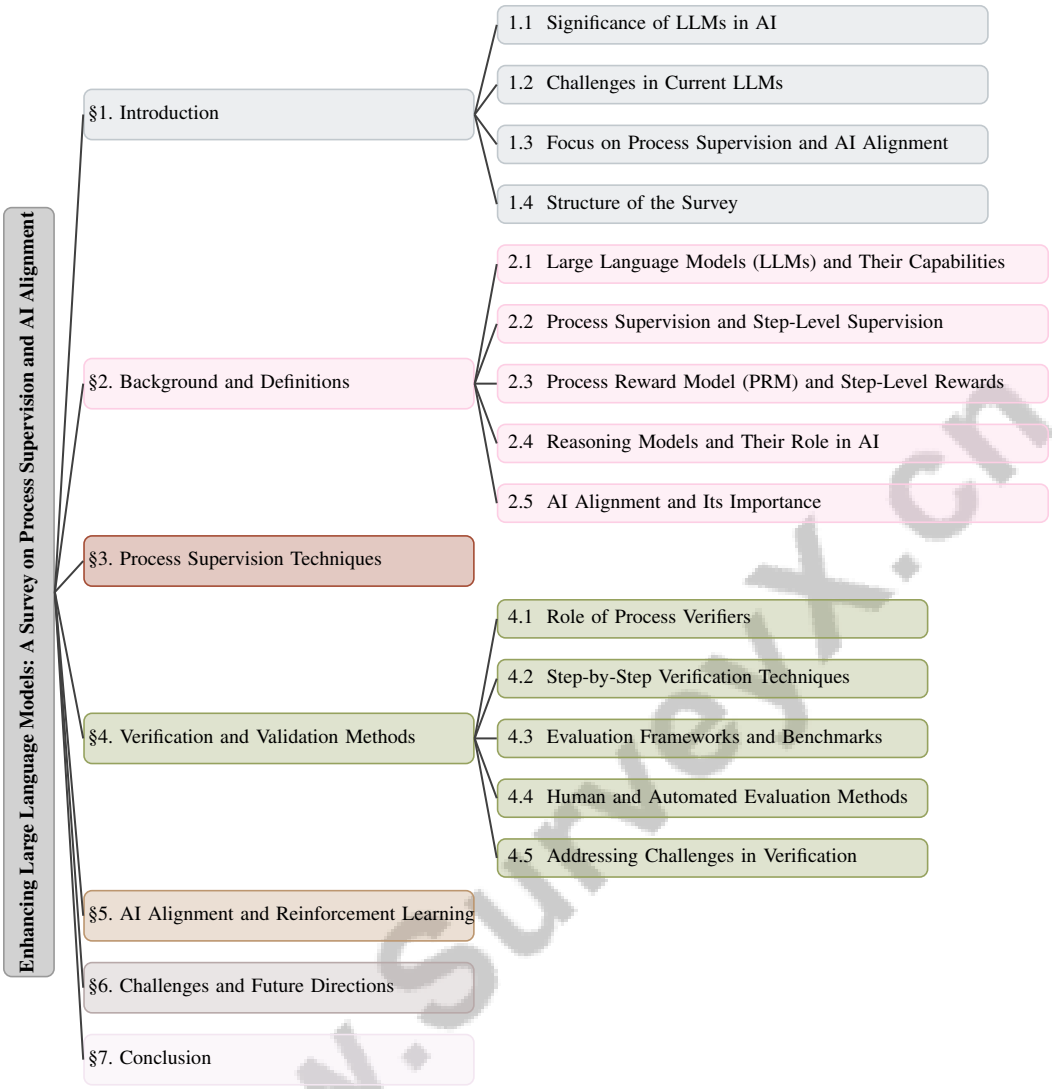


Figure 1: chapter structure

1.2 Challenges in Current LLMs

LLMs encounter various challenges that hinder their effectiveness across applications. A primary issue is the insufficient exploration of memory mechanisms, which limits the development of comprehensive theoretical frameworks for understanding memory's role in LLMs [11]. This shortfall restricts their capacity to capture nuanced, context-specific knowledge essential for predictive analytics and decision-making [12].

Methodological inefficiencies, such as those seen in Graph2Text for graph reasoning tasks, further complicate their performance, as these models struggle to derive graph structures from extensive text [13]. The absence of informative benchmarks to accurately assess LLM capabilities exacerbates these issues [7]. Additionally, LLMs often misinterpret user intents due to the inherent ambiguity in natural language, complicating tasks like visualization generation [3].

In STEM fields, LLMs exhibit limited performance in complex reasoning tasks, particularly in multi-step problem-solving and deeper comprehension [14]. Ambiguity in action descriptions and slow inference speeds further impede real-time decision-making [6]. Vulnerabilities to prompt hacking attacks can lead to harmful or biased outputs [15], while difficulties with logical reasoning and identifying fallacies persist [16]. Class imbalances and complex multi-label classification in public affairs present additional challenges [2].

AI-generated, for reference only.

Concerns about privacy and security arise as LLMs may memorize sensitive information from training data, posing risks in applications [17]. Existing benchmarks also struggle with spam tactics' concept drift, requiring costly updates and large labeled datasets [5]. Moreover, LLMs often rely on fixed knowledge, leading to inaccuracies known as hallucinations [8]. Fine-tuning LLMs in low-data regimes remains challenging, as traditional data augmentation methods frequently fall short [18]. The integration of LLMs with ASR systems presents challenges such as acoustic information loss that compounds recognition errors [4].

The resource demands of LLMs contribute to hallucination phenomena, impacting keyword extraction accuracy [1]. The unclear theoretical basis for attributing credences to LLMs raises questions regarding their possession of mental states justifying such attributions [19]. Collecting high-quality human feedback is challenging due to significant costs and difficulties in accurately simulating human preferences [9]. Finally, existing methods often fail to leverage extensive natural language data and images effectively, limiting personalized and contextually relevant recommendations [20]. Demographic biases and inaccuracies in LLM annotations further undermine their reliability and applicability [10]. These multifaceted challenges underscore the necessity for ongoing research to enhance LLM robustness and applicability across diverse fields.

1.3 Focus on Process Supervision and AI Alignment

Enhancing LLMs through process supervision and AI alignment is vital for addressing technical challenges and ensuring alignment with human values. Process supervision frameworks improve LLM interpretability and performance, enabling more precise outputs. For instance, step-level supervision in systematic reasoning pipelines, such as those used in visualization generation, exemplifies the structured approach necessary for optimizing model outputs [3]. The application of Reinforcement Learning with Minimum Editing Constraint (RLMEC) provides essential fine-grained supervision for complex reasoning tasks [21].

AI alignment ensures that LLMs operate in accordance with human ethical standards, addressing the misalignment between outputs and human intentions, often stemming from pre-training on noisy data [9]. The exploration of LLMs' moral compass in ethical decision-making scenarios emphasizes the importance of alignment in AI development, highlighting the need for models that adhere to societal norms [22].

Integrating multi-modal information processing through LLMs into recommender systems enhances performance by leveraging diverse data sources, thereby improving adaptability and utility across applications [20]. Additionally, confidence-driven inference methods that combine LLM annotations with human inputs signify a substantial advancement in enhancing output reliability and accuracy [10].

This survey focuses on process supervision and AI alignment to improve LLM reliability, interpretability, and ethical alignment, addressing existing challenges and promoting their application across various fields. These efforts are crucial for ensuring that LLMs meet technical performance standards while adhering to ethical and societal expectations, fostering trust and reliability in AI systems. Furthermore, interpreting LLM credence attributions as literal expressions of beliefs about LLM credences highlights the need for a deeper understanding of the cognitive frameworks that underpin these models [19].

1.4 Structure of the Survey

This survey systematically explores the enhancement of LLMs through process supervision and AI alignment, addressing technical challenges and ethical considerations. It begins with an introduction to the significance of LLMs in AI, outlining their capabilities and the challenges they face. The subsequent section emphasizes the importance of process supervision and AI alignment in improving LLM interpretability and ensuring ethical alignment with human values. This focus addresses ethical challenges unique to LLMs, including hallucination, accountability, and bias, while advocating for transparency and interdisciplinary collaboration. Developing tailored ethical frameworks and dynamic auditing systems aims to foster responsible LLM development that prioritizes human understanding and societal impacts [23, 12, 24].

The second section provides essential background and definitions, detailing key concepts such as LLM capabilities, process supervision, step-level supervision, and AI alignment's importance. This foundation sets the stage for a deeper exploration of process supervision techniques in the third section, where methodologies like prompt engineering, semantic routing, and data augmentation are discussed.

Verification and validation methods form the core of the fourth section, emphasizing process verifiers and step-by-step verification techniques to ensure LLM accuracy and reliability. The fifth section delves into AI alignment and reinforcement learning, exploring how these frameworks contribute to aligning LLMs with human values and ethical standards, as exemplified by the development of 'Arithmetic-GPT' for enhanced decision-making alignment [25].

The survey concludes with a comprehensive discussion of current challenges and future directions for LLMs, highlighting technological advancements and potential research avenues. It identifies areas for improvement, such as enhancing table-to-text generation capabilities for real-world applications, addressing ethical complexities like accountability and bias, and optimizing post-ranking processes in search engines. Emphasizing interdisciplinary collaboration and tailored ethical frameworks aims to guide responsible LLM integration, ultimately enhancing their effectiveness and reliability across applications, including information retrieval and systematic reviews [26, 27, 28, 24, 29]. This structure provides a holistic understanding of the current landscape and future prospects in LLM development, ensuring that these models achieve technical excellence while adhering to ethical and societal expectations.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Large Language Models (LLMs) and Their Capabilities

Large Language Models (LLMs) are integral to artificial intelligence, excelling in understanding and generating human-like text across diverse applications. Built on the Transformer architecture, they exist in encoder-only, decoder-only, and encoder-decoder configurations, enabling complex linguistic tasks [7]. Their structured framework enhances reasoning, comprehension, and core language modeling capabilities [7].

LLMs excel in natural language processing, significantly aiding keyword extraction from scientific literature [1] and translating qualitative insights into quantifiable features for predictive analytics [12]. Tools like ChartGPT demonstrate their ability to generate visual representations from abstract language inputs [3].

Integration with graph learning models, such as GraphLLM, enhances reasoning on graph data for sophisticated analytics [13], benefiting applications like public affairs document classification [2]. LLMs also improve Automatic Speech Recognition (ASR) by aligning auditory and textual information [4].

In multi-modal environments, LLMs leverage deep learning to integrate diverse data types, enhancing recommendation accuracy [20]. Their adaptability is further evidenced in zero-shot spam email classification, achieving performance benchmarks without extensive retraining [5].

LLMs contribute to embodied AI, as seen in models like LARM, predicting actions based on text and multi-view images [6]. Reward-guided tree search frameworks, such as STILL-1, enhance reasoning by integrating policy models, reward models, and search algorithms [14].

Despite their capabilities, challenges remain in ensuring LLM reliability and ethical alignment, particularly in data annotation tasks where reliability is crucial [10]. Understanding LLMs' credences as mental states reflecting confidence is essential for grasping their capabilities and limitations [19]. Ongoing refinement is necessary for achieving technical and ethical standards across domains.

2.2 Process Supervision and Step-Level Supervision

Process and step-level supervision are critical for enhancing LLM reasoning capabilities through structured guidance and feedback. These methodologies improve the accuracy and relevance of LLM outputs by guiding reasoning processes. Process supervision, involving domain-specific data training,

enhances tasks like keyword extraction via prompt engineering [1], and proves useful in zero-shot spam email classification without extensive fine-tuning [5].

Innovative approaches address current limitations, such as the Speech Foundation Encoder-LLM ASR (SFE-LLM-ASR) method, which enhances ASR performance by aligning features from speech and text modalities [4]. Multi-modal recommendation systems unify modalities in latent space, enhancing discriminative power [20].

Step-level supervision optimizes individual reasoning steps, providing rich feedback for complex tasks [21]. Confidence-driven inference methods, utilizing human annotations based on LLM confidence scores, improve statistical estimates and output reliability [10]. These methodologies pave the way for robust LLMs capable of intricate reasoning with precision and contextual understanding, leveraging LLMs' inner states for factual detection and performance improvement.

2.3 Process Reward Model (PRM) and Step-Level Rewards

Process Reward Models (PRMs) and step-level rewards optimize LLM performance by providing structured feedback that enhances decision-making and reasoning. PRMs improve reasoning pathways with dense feedback, enhancing learning efficiency and accuracy in domain-specific applications like code generation [30].

Reward functions based on the Value of Computation (VOC) penalize unnecessary reasoning steps, enabling efficient reasoning processes while maintaining performance [31]. Step-level rewards focus on individual reasoning steps, facilitating detailed training processes with precision and contextual understanding [32].

These methodologies are employed across domains, including medical diagnostics, ensuring LLM outputs are technically sound and contextually relevant [33]. Tree search algorithms like REward BAlanced SEarch (REBASE) achieve Pareto-optimal trade-offs, illustrating PRMs' utility in optimizing reasoning pathways [34].

The evolution of these methodologies, including Iterative Prompting Optimization (IPO), highlights LLM optimization's dynamic nature [35]. Benchmarks in public affairs multi-label classification [2] and zero-shot spam email classification [5] emphasize PRMs and step-level rewards' importance in optimizing LLM capabilities.

2.4 Reasoning Models and Their Role in AI

Reasoning models are fundamental in LLM architecture, providing structured frameworks for logical reasoning and problem-solving. They enhance cognitive capabilities, emulating human-like reasoning critical for decision-making and complex tasks. Theoretical foundations draw from cognitive psychology and linguistics, illuminating LLM language processing and generation [36].

Cognitive science perspectives, employing factor analysis, highlight the interplay between language comprehension and reasoning [7]. This basis is essential for developing reasoning models capable of interpreting complex linguistic inputs.

Reasoning models enhance LLM self-verification abilities, allowing them to identify and rectify logical fallacies [16]. This capability ensures output reliability and accuracy, especially in tasks demanding rigorous analysis.

Integration of visual and textual information, as in models like LARM, maintains high inference speeds and informed action predictions [6]. This is beneficial in rapid decision-making environments, showcasing reasoning models' versatility.

Reasoning models optimize market dynamics through frameworks like double auction mechanisms, achieving market equilibrium and welfare maximization [37]. These models underscore structured reasoning's significance beyond traditional AI domains.

Current LLM limitations in logical reasoning, particularly in commonsense, numerical, and qualitative reasoning, are revealed through evaluations like LogicBench. Innovative approaches, such as Chain-of-Thought and Program-of-Thought, improve reasoning capabilities and facilitate self-verification processes, enabling LLMs to learn from errors effectively [16, 26, 38, 39]. By integrating structured

reasoning frameworks, LLMs achieve technical excellence while adhering to ethical standards, fostering trust in AI systems.

2.5 AI Alignment and Its Importance

AI alignment is crucial for developing LLMs, ensuring models operate according to human values and ethical standards. This alignment mitigates ethical dilemmas, preventing outcomes conflicting with societal norms [9]. AI alignment involves tailoring guidelines to specific inputs, enhancing output precision through methods like Reinforcement Learning with Minimum Editing Constraint (RLMEC) for error correction [21].

Ensuring fairness in natural language processing systems is a challenge in AI alignment, as biases against marginalized groups must be addressed to foster trust and reliability [9]. AI safety is underscored by preventing catastrophic responses, highlighting robustness and applicability concerns [8].

Evaluating LLM performance in providing reliable advice is crucial in sensitive domains like finance or law, emphasizing robust alignment frameworks for accurate, value-aligned outputs. Perspectives on LLM credences raise questions about attribution methods' validity [19].

AI alignment addresses robustness issues in model editing, essential for safe deployment in high-stakes domains [17]. Fostering trust in AI systems requires considering reliability, openness, and task characteristics for effective interactions.

AI alignment ensures LLMs achieve technical excellence while adhering to ethical standards, harnessing their potential across sectors. Addressing ethical challenges like privacy, fairness, and accountability is crucial, with transparency measures mitigating AI deployment risks. Integrating human insights and ethical frameworks tailored to domains establishes a robust foundation for responsible AI advancement, enhancing societal impact [23, 12, 10, 24].

3 Process Supervision Techniques

Category	Feature	Method
Novel Prompt Engineering Techniques	Domain and Task Specific	LLM-FE[12], Path-LLaMA[40], LLMCRS[41]
	Error and Component Refinement	MEMT[42], MRKL[43]
	Structured Reasoning and Exploration	STILL-1[14], CG[3]
Semantic Routing and Intent Extraction	Confidence-Based Techniques	CDI[10]
Data Augmentation and Logical Reasoning	Logical Reasoning Enhancement	CoT[44], RFT[45], SIPF[46]
	Efficiency and Accuracy Improvement	VQ[47]
Automated Process Supervision Data Collection	Multi-Modal Data Handling	LARM[6]

Table 1: This table provides a comprehensive overview of various methodologies used to enhance the performance and reliability of Large Language Models (LLMs) through process supervision techniques. It categorizes these methodologies into key areas such as novel prompt engineering techniques, semantic routing and intent extraction, data augmentation and logical reasoning, and automated process supervision data collection, detailing specific features and methods within each category. The table serves as a valuable resource for understanding the diverse approaches employed to optimize LLM capabilities across different applications.

Innovative process supervision methodologies have significantly enhanced the reasoning capabilities of Large Language Models (LLMs). Automated process supervision assigns intermediate rewards during reasoning, addressing limitations of traditional methods that only evaluate final outcomes. The OmegaPRM algorithm exemplifies this by efficiently collecting high-quality process supervision data, improving mathematical reasoning performance. Consequently, instruction-tuned LLMs exhibit increased success rates on complex multi-step reasoning tasks, demonstrating the effectiveness of process supervision in enhancing LLM capabilities [48, 26, 49]. These techniques enable efficient handling of complex tasks and adaptation to diverse applications. Table 1 offers a detailed classification of process supervision methodologies, highlighting their contributions to advancing Large Language Model performance and adaptability. Table 3 presents a comparative analysis of various process supervision methodologies, underscoring their respective roles in optimizing Large Language Model performance and adaptability. To further elucidate the intricacies of these methodologies, **??** presents a hierarchical structure of process supervision techniques. This figure highlights key cate-

gories such as novel prompt engineering, semantic routing and intent extraction, data augmentation and logical reasoning, and automated process supervision data collection. Each category is further divided into subcategories that detail specific methodologies and their contributions to enhancing LLM performance and reliability. We first explore novel prompt engineering techniques, foundational to optimizing LLM performance and adaptability.

3.1 Novel Prompt Engineering Techniques

Method Name	Methodological Structure	Domain-Specific Applications	Reasoning Enhancement
Path-LLaMA[40]	Task Decomposition	Pathology Report Extraction	Multi-layered Networks
LLMCRS[41]	Structured Workflow	Conversational Recommendations	Expert Models
MEMT[42]	Model Editing Techniques	Machine Translation Refinement	Refining Located Components
MRKL[43]	Modular Architecture	Knowledge Tasks	Discrete Reasoning Modules
STILL-1[14]	Reward-guided Tree	-	Guided Exploration
LLM-FE[12]	Llm-based Feature Encoding	Transaction Data Extraction	Multi-layered Neural Networks
CG[3]	Step-by-step Reasoning	Visualization Generation Tasks	Multi-layered Neural Networks
CDI[10]	Task Decomposition	-	Guided Exploration
CoT[44]	Decomposing Tasks	Zero-shot Spam	Multi-layered Neural

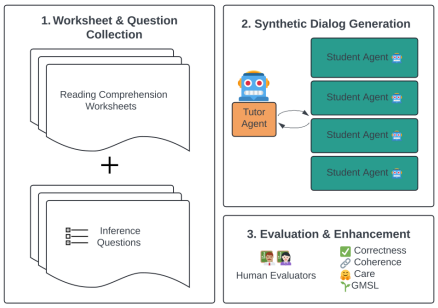
Table 2: This table provides a comprehensive overview of various novel prompt engineering techniques, detailing their methodological structures, domain-specific applications, and reasoning enhancements. Each method is designed to improve the performance and adaptability of large language models (LLMs) across diverse tasks, highlighting the importance of structured approaches in enhancing LLM outputs.

Table 2 outlines the diverse range of novel prompt engineering techniques that have been developed to enhance the performance of large language models (LLMs) in various applications. Novel prompt engineering techniques advance LLM performance and adaptability through structured methodologies that enhance outputs across various applications. The Path-LLaMA method exemplifies this by tailoring prompt engineering for local LLMs, improving structured data extraction from pathology reports and demonstrating the importance of domain-specific prompt engineering for data extraction accuracy [40]. Frameworks like LLMCRS refine LLM capabilities by decomposing complex tasks into sub-tasks, increasing output precision and effectiveness [41]. Methodologies that manipulate error-prone components in translation quality enhancement showcase the strategic use of prompt engineering to address specific language processing challenges [42]. Geometric features derived from LLMs' intrinsic dimensions present a novel approach to augmenting toxicity detection without extensive labeled datasets [50].

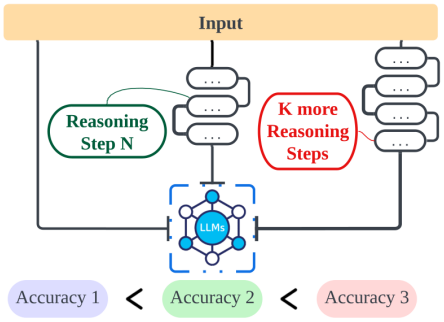
The MRKL system employs a flexible architecture to intelligently route inputs among different modules, enhancing reasoning and knowledge integration [43]. Similarly, the STILL-1 framework integrates a policy model, a reward model, and a search algorithm to enhance LLM reasoning through guided exploration [14]. Practical applications include using transaction data to guide LLMs in extracting relevant features indicative of complex activities, translating expert intuition into quantifiable insights [12]. The ChartGPT methodology employs a step-by-step reasoning approach to improve chart generation accuracy [3]. The CONFIDENCE-DRIVEN INFERENCE method enhances annotation processes by integrating LLM annotations and confidence scores [10]. Prompt design is crucial for optimizing LLMs in tasks like zero-shot spam classification, where design significantly impacts model performance [5].

Collectively, these novel prompt engineering techniques significantly advance LLMs, ensuring robust application across diverse environments while maintaining high ethical and societal standards. By harnessing advancements such as innovative table-to-text generation capabilities and the LLM2LLM iterative data enhancement strategy, models effectively tackle complex challenges, enhancing user efficiency and reducing the need for extensive data curation [18, 29].

As depicted in Figure 2, advanced systems for enhancing cognitive skills and reasoning capabilities are crucial. The "Synthetic Tutoring System for Reading Comprehension" and "Deep Reasoning with Multiple Reasoning Steps" illustrate innovative approaches to these challenges. The Synthetic Tutoring System enhances reading comprehension through a structured framework comprising worksheet creation, synthetic dialog generation, and iterative evaluation, emphasizing organized question collection and inferential reasoning. Meanwhile, the Deep Reasoning example employs a sophisticated flowchart model with layered neural networks to process inputs and produce outputs with varying accuracy, underscoring the significance of multi-layered reasoning in achieving higher



(a) Synthetic Tutoring System for Reading Comprehension[51]



(b) Deep Reasoning with Multiple Reasoning Steps[44]

Figure 2: Examples of Novel Prompt Engineering Techniques

precision. Together, these examples highlight the potential of novel prompt engineering techniques in advancing educational and reasoning processes [51, 44].

3.2 Semantic Routing and Intent Extraction

Semantic routing and intent extraction are essential in LLM supervision, enhancing their ability to accurately interpret user inputs and generate contextually appropriate outputs. By implementing semantic routing, LLMs leverage structured prompts to optimize intent-based networking, particularly in applications like 5G core network management. This approach improves deployment efficiency and accuracy, facilitating automation of complex tasks such as data extraction and keyword identification [52, 48, 26, 1]. These methodologies enhance LLM performance by ensuring inputs are routed to appropriate processing modules and user intents are accurately extracted.

Modular architectures, like those in systems such as MRKL, significantly enhance semantic routing by enabling intelligent input routing across various modules, optimizing reasoning and knowledge integration [10]. Intent extraction involves accurately identifying and interpreting user intents from natural language inputs, crucial for applications requiring nuanced understanding and response generation. Techniques for intent extraction often utilize confidence-driven inference methods, combining LLM annotations with confidence scores to enhance reliability and accuracy [10]. By prioritizing accurate interpretations based on confidence scores, these methods improve the overall quality of LLM outputs.

The synergy of semantic routing and intent extraction is vital for optimizing process supervision in LLMs, enabling precise and contextually appropriate responses. Continuous refinement of operational processes enhances LLM performance across diverse applications, boosting accuracy and reliability while automating various task stages, such as literature reviews, data extraction, and content generation. Recent studies highlight the superior capabilities of GPT-based models in data extraction, achieving a mean precision of 83.0

3.3 Data Augmentation and Logical Reasoning

Data augmentation techniques are pivotal in enhancing LLM logical reasoning capabilities by providing enriched datasets for comprehensive training and evaluation. The Rejection Sampling Fine-Tuning (RFT) method generates and collects correct reasoning paths as augmented datasets to improve model performance [45]. This method exposes LLMs to diverse reasoning patterns, enhancing their generalization across different logical scenarios.

Resources like LogicBench, which includes 25 unique reasoning patterns focusing on single inference rules, advance the systematic evaluation of logical reasoning [38]. Techniques that manipulate reasoning steps within prompts are crucial for understanding their impact on LLM performance. Methods that expand or compress reasoning steps while controlling for other variables provide valuable insights into how different reasoning structures affect model outputs [44]. Preference optimization methods,

which involve sampling reasoning paths and scoring them based on intermediate correctness, further enhance LLM performance [46].

VerifierQ computes Q-values in parallel for multiple steps, enhancing training efficiency and accuracy in evaluating the correctness of generated solutions [47]. Data augmentation techniques, particularly when integrated with advanced logical reasoning methodologies, play a critical role in enhancing LLM performance. Recent innovations, such as the LLM2LLM strategy, allow for iterative augmentation of data by utilizing a teacher LLM to generate synthetic examples based on previously misclassified instances, improving model accuracy in low-data scenarios. Methods like hashing bias-inducing words reduce cognitive biases, further refining LLM performance [53, 54, 55, 14, 18].

3.4 Automated Process Supervision Data Collection

Automated process supervision data collection is crucial for optimizing LLM training and evaluation by systematically curating high-quality datasets that enhance model performance and reliability. Techniques such as integrating multi-modal inputs, as demonstrated by LARM, enable LLMs to generate precise action predictions in real-time, enhancing their ability to interact effectively in complex environments [6]. Evaluating LLM credences through experimental techniques, including prompting models to report their confidence levels and consistency-based estimation, refines the data collection process [19]. By measuring output probabilities and ensuring consistency in responses, these techniques contribute to developing robust data collection frameworks.

Advanced methodologies focus on enhancing data collection integrity by implementing safeguards against adversarial influences, ensuring data accuracy and reliability amid potential manipulation or misinformation. This focus addresses vulnerabilities identified in recent studies, which reveal that even advanced models can inadvertently propagate errors or misconceptions when providing security and privacy advice. By leveraging improved predictive analytics and knowledge-driven frameworks, these methodologies aim to integrate expert insights with machine learning techniques, fortifying data collection against adversarial threats while maintaining quality and authenticity [56, 57, 12, 58]. Techniques that protect against data poisoning attacks ensure the integrity of training datasets, maintaining the reliability and ethical alignment of LLM outputs. Structured methodologies, such as decomposing complex questions into subquestions and sequentially addressing them, exemplify strategic approaches in data collection to enhance model reasoning capabilities.

Incorporating task-specific instructions and filtering uncertain predictions improve the quality of collected data, ensuring LLMs are trained on datasets that are both comprehensive and contextually relevant. These methodologies collectively advance automated process supervision data collection, facilitating efficient extraction of relevant information and supporting ongoing refinement and development of LLMs. By employing innovative methodologies such as systematically converting investigator insights into quantifiable features, enhancing table-to-text generation capabilities, and bias-reduction techniques, the data collection process for LLMs becomes significantly more efficient. These advancements improve models' reasoning capabilities and contextual understanding, facilitating the integration of expert knowledge and enhancing decision-making accuracy across various applications [59, 12, 53, 29].

Feature	Novel Prompt Engineering Techniques	Semantic Routing and Intent Extraction	Data Augmentation and Logical Reasoning
Optimization Focus	Prompt Design	Input Routing	Dataset Enrichment
Key Technique	Structured Methodologies	Modular Architectures	Rejection Sampling
Performance Impact	Improves Adaptability	Enhances Interpretation	Improves Reasoning

Table 3: Comparison of key methodologies in process supervision for Large Language Models, focusing on novel prompt engineering techniques, semantic routing and intent extraction, and data augmentation with logical reasoning. Each method is evaluated based on its optimization focus, key technique, and performance impact, highlighting their contributions to enhancing adaptability, interpretation, and reasoning capabilities.

4 Verification and Validation Methods

Understanding verification and validation methods is crucial for ensuring the reliability and accuracy of Large Language Models (LLMs). This section examines the integral role of process verifiers in

enhancing model performance and ethical alignment, as well as their significance in maintaining the integrity of LLM outputs across various applications.

4.1 Role of Process Verifiers

Process verifiers are pivotal in ensuring the accuracy, reliability, and ethical alignment of LLM outputs. These mechanisms assess and validate LLM reasoning processes, thus enhancing performance and trustworthiness. The integration of process verifiers, such as those in LLMCRS, is vital for maintaining response fidelity, ensuring outputs are accurate and contextually appropriate [41]. In toxicity detection, geometric features from LLMs surpass existing methods, showcasing verifiers' potential to leverage intrinsic features for mitigating harmful content [50]. GraphLLM's simultaneous processing of node information and graph structure exemplifies verifiers' role in enhancing reasoning capabilities in graph tasks [13].

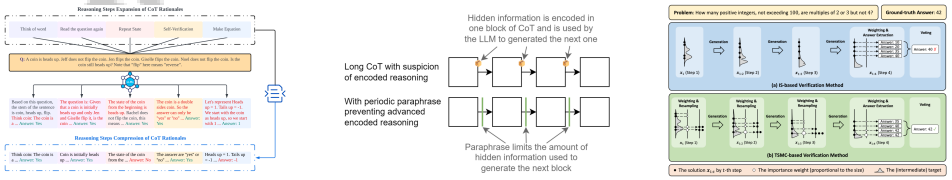
In practice, process verifiers reduce the need for human annotations while maintaining statistical validity, crucial for large-scale deployments [10]. Techniques like RLMEC provide fine-grained supervision, refining reasoning capabilities by meticulously evaluating each step [21]. Evaluation frameworks, including self-verification benchmarks, deepen understanding of LLM reasoning and verification abilities [16]. Employing trained verifiers for intermediate reasoning steps improves problem-solving, especially in mathematics and coding. Techniques like Model-induced Process Supervision (MiPS) automate data curation, boosting reasoning accuracy, while VerifierQ optimizes verification through reinforcement learning, enhancing performance on benchmark tasks and addressing LLMs' reasoning limitations [47, 39, 60].

4.2 Step-by-Step Verification Techniques

Step-by-step verification techniques enhance LLM reliability by systematically evaluating each reasoning stage. These techniques decompose complex tasks into smaller components for detailed analysis. Step-level supervision in reasoning pipelines, such as those used in visualization generation, exemplifies the structured approach needed for improved outputs [3].

VerifierQ, which computes Q-values in parallel, enhances training efficiency and solution correctness evaluation, ensuring each step contributes meaningfully [47]. Preference optimization methods, which sample reasoning paths and score them based on intermediate correctness, refine LLM performance [46]. Frameworks like LogicBench, focusing on single inference rules, provide structured assessments of logical reasoning strengths and weaknesses, guiding robust LLM development [38].

Overall, step-by-step verification techniques are indispensable for robust LLM development, ensuring accuracy and reliability in complex reasoning tasks. By incorporating structured verification frameworks, LLMs can enhance technical performance, ensuring compliance with ethical standards. This integration improves reasoning and verification abilities, promoting accountability and transparency, fostering trust in AI systems, and addressing challenges like error detection and bias reduction [61, 16, 62, 39, 24].



(a) Reasoning Steps Expansion of CoT Rationales[44]

(b) Long CoT with suspicion of encoded reasoning and periodic paraphrase preventing advanced encoded reasoning[63]

(c) Verification of a Problem with a Ground-Truth Answer[64]

Figure 3: Examples of Step-by-Step Verification Techniques

As shown in Figure 3, step-by-step verification techniques illustrate meticulous processes for ensuring computational reasoning accuracy. The figures depict methodologies like "Reasoning Steps Expansion of CoT Rationales," highlighting the importance of segmenting complex reasoning. "Long CoT with suspicion of encoded reasoning" emphasizes transparency, while "Verification of a Problem with a

Ground-Truth Answer" underscores ground-truth's role in validating computational solutions. These examples highlight the significance of step-by-step techniques in enhancing reasoning and validation robustness [44, 63, 64].

4.3 Evaluation Frameworks and Benchmarks

Benchmark	Size	Domain	Task Format	Metric
LLM-TC[2]	33,147	Public Affairs	Multi-label Classification	True Positive Rate, True Negative Rate
ClashEval[58]	1,294	Healthcare	Question Answering	Accuracy, Prior Bias
LLM-RB[65]	100,000	Natural Language Inference	Multiple-choice Question Answering	Accuracy, ROUGE
LLM-IS[66]	1,500	Conversational AI	Language Style Imitation	Human Evaluation, Automatic Evaluation
LogicBench[38]	12,908	Logical Reasoning	Binary Question-Answering (bqa) And Multiple-Choice Question-Answering (mcqa)	Accuracy
GSM8K[49]	8,000	Mathematics	Question Answering	Final-answer error rate, Trace error rate
TTD[51]	1,000	Reading Comprehension	Dialog-based Tutoring	Helpfulness, Success@k
LLMs4OL[67]	3,000,000	Biomedical	Term Typing	MAP@1, F1-score

Table 4: This table presents a comprehensive overview of various benchmarks used for evaluating large language models (LLMs) across multiple domains. Each benchmark is characterized by its size, specific domain of application, task format, and the metrics employed for performance evaluation. The diversity in benchmarks underscores the importance of tailored evaluation frameworks to address the unique challenges and requirements of different application areas.

Evaluation frameworks and benchmarks are vital for systematically assessing LLMs, providing structured methodologies to measure performance across tasks and domains. These benchmarks highlight LLMs' generalization to real-world applications but require nuanced approaches due to varied test prompt distributions. Continuous refinement in evaluation methods is crucial to keep pace with LLM advancements [1, 65, 26]. Table 4 provides a detailed overview of representative benchmarks employed in the evaluation of large language models, highlighting the diversity in domains, task formats, and evaluation metrics.

Metrics like True Positive Rate (TPR) and True Negative Rate (TNR) address class imbalances in tasks like public affairs document classification, enhancing LLM applicability and reliability [2]. In automatic speech recognition (ASR), evaluations using extensive datasets ensure LLMs handle diverse speech data, improving real-world adaptability [4]. Benchmarks for implicit classification tasks employ metrics like ROC curve, AUC, and F1 score, providing a comprehensive assessment of LLM capabilities, particularly in complex linguistic inputs [26, 68, 69, 12, 70].

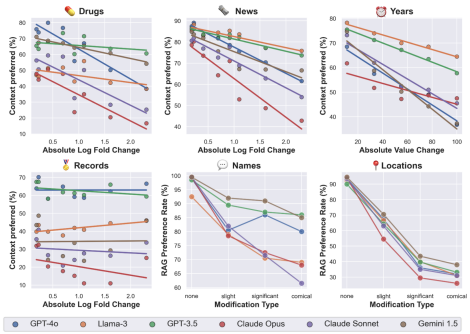
Evaluation frameworks and benchmarks are integral to LLM development, offering insights into capabilities and guiding ongoing advancements. Structured evaluation methodologies address ethical challenges like hallucination and accountability, enhancing AI system transparency and reliability, aligning outputs with societal values and ethical standards [71, 61, 23, 62, 24].

As shown in Figure 4, the domain of verification and validation methods, alongside evaluation frameworks and benchmarks, is exemplified by two illustrations. "Comparison of Context Preference Rates Across Different Textual Data Types and Metrics" shows a comparative analysis of context preference rates, while "LLM Ranking: A Visual Explanation" visually depicts the LLM ranking process. These examples underscore robust evaluation frameworks and benchmarks' importance in LLM development [58, 65].

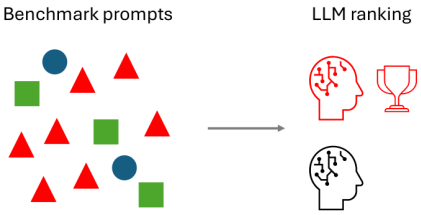
4.4 Human and Automated Evaluation Methods

Evaluating LLMs requires a comprehensive approach encompassing human and automated methods to ensure accuracy, reliability, and ethical alignment. Human evaluation involves expert assessments, providing qualitative feedback on contextual relevance and ethical implications, crucial in domains like medical diagnostics or legal advice [33]. Human evaluations identify subtle biases and ethical concerns that automated methods might overlook, ensuring adherence to societal and ethical standards.

Automated evaluation methods use quantitative metrics and algorithmic assessments to evaluate LLM performance. Metrics like accuracy, precision, recall, and F1 score provide objective performance



(a) Comparison of Context Preference Rates Across Different Textual Data Types and Metrics[58]



(b) LLM Ranking: A Visual Explanation[65]

Figure 4: Examples of Evaluation Frameworks and Benchmarks

measures. Automated methods excel in large-scale assessments where data volume necessitates efficiency. ROC curves and AUC metrics in implicit classification tasks offer robust frameworks for evaluating classification capabilities [5].

Integrating human and automated methods ensures comprehensive LLM assessment, identifying strengths and weaknesses, leading to robust evaluations across applications. While automated methods offer scalability, human evaluations provide insights into ethical dimensions. This dual approach ensures LLMs achieve technical excellence and align with ethical expectations, fostering trust in AI systems [70, 26].

4.5 Addressing Challenges in Verification

Verification of LLMs presents challenges due to task complexity and variability. A significant challenge is the subjective nature of harm detection, complicating consistent verification framework development. This subjectivity arises from diverse contexts where harm interpretation varies [72].

Data annotation for verification is challenging, requiring consistent labeling across scenarios to ensure reliable harm identification and mitigation. This uniformity is difficult due to human judgment variability and language complexity [72].

Addressing these challenges requires robust verification frameworks incorporating automated and human evaluation methods. Automated methods provide scalable solutions, while human evaluations offer nuanced insights into ethical implications. Integrating advanced verification strategies, like collaborative reasoning and specialized detector models, enhances verification processes, ensuring LLMs operate safely across applications. This approach improves output accuracy and reliability, addressing LLM risks such as biased outputs, fostering trust and efficiency [73, 72, 12, 39].

Adaptive learning techniques, allowing LLMs to adjust outputs based on real-time feedback, further improve verification. These techniques empower LLMs to analyze mistakes and enhance output quality. A two-stage approach combining safety-trained models with guideline libraries identifies risks and generates value-aligned responses, improving reliability. Document-wise memory architecture and self-correction strategies refine contextually relevant content production, aligning outputs with ethical standards and user expectations [71, 10, 74, 75, 58]. Continuous refinement and comprehensive verification strategies effectively mitigate LLM verification challenges, ensuring safe deployment.

5 AI Alignment and Reinforcement Learning

5.1 Conceptual Framework of AI Alignment

The conceptual framework of AI alignment is pivotal for ensuring that Large Language Models (LLMs) adhere to human values and ethical standards. This framework encompasses methodologies that enhance reasoning capabilities and align outputs with human preferences, exemplified by the

LSC-alignment method, which uses a log-sigmoid transformation to align LLM outputs with user expectations [76]. Improving evaluation methods is essential for advancing LLM performance in mathematical reasoning, as demonstrated in studies of LLM mastery over mathematical tasks [28]. Ethical reasoning frameworks further refine this alignment by categorizing LLMs according to ethical perspectives, such as utilitarian and values-based ethics, crucial for models adhering to ethical norms [22].

Personalization plays a significant role in alignment, as evidenced by the PORTLLM approach, which incorporates personalized knowledge from previous fine-tuning into new model versions, enhancing performance without further training [77]. Local LLMs tailored to specific domains, like healthcare, ensure alignment with domain-specific requirements [40]. Graph reasoning is also critical, with models like GraphLLM improving graph reasoning accuracy while minimizing context length, underscoring the need for task-specific alignment [13]. Future research aims to expand model capabilities for broader visualizations, aligning LLMs more closely with user needs [3].

This interdisciplinary framework emphasizes transparency and human-centered approaches, addressing ethical challenges such as hallucination and accountability. It advocates for tailored ethical frameworks and dynamic auditing systems to mitigate risks and promote responsible LLM development [23, 24]. By integrating process-based feedback, personalized knowledge, and ethical reasoning strategies, this framework fosters robust AI systems aligned with societal expectations.

5.2 Role of Reinforcement Learning in AI Alignment

Reinforcement learning (RL) is crucial for aligning Large Language Models (LLMs) with human values, optimizing decision-making through structured feedback. The LSC-alignment method exemplifies RL's application in AI alignment, utilizing transformed rewards to guide LLMs towards preferred outputs [76]. Recent advancements, like step-wise preference optimization techniques such as Step-DPO, provide fine-grained supervision to enhance alignment with human preferences in complex tasks [78].

Model surgery techniques enable significant modulation of LLM behavior with minimal computational resources, addressing concerns related to toxicity and jailbreak scenarios without extensive fine-tuning, showcasing RL's potential in aligning LLM outputs with ethical standards [79]. Integrating Bayesian reward models, which incorporate uncertainty, enhances the evaluation of LLM responses, ensuring alignment with human preferences [80]. Automated feedback strategies further illustrate RL's utility in refining LLM performance [81]. The AutoFlow framework, for instance, exemplifies workflow optimization through RL, using performance evaluations as rewards to iteratively refine LLM outputs [82].

Attention steering methods direct LLMs to focus on relevant information, enhancing output precision [83]. Benchmarks are essential for selecting appropriate models and formal languages for knowledge base question answering (KBQA) tasks, improving LLM effectiveness in reasoning [84]. RL is integral to aligning LLMs with human values, optimizing decision-making processes, and ensuring ethically and contextually aligned outputs. By incorporating RL techniques, LLMs enhance predictive analytics capabilities, integrating expert domain knowledge into quantifiable features, thereby improving accuracy and fostering trust [10, 12, 85].

5.3 Integration of External Knowledge and Verification

Integrating external knowledge into Large Language Models (LLMs) is crucial for enhancing AI alignment and verification processes, improving the accuracy and contextual relevance of model outputs. The LLMs4OM framework illustrates that LLMs can facilitate ontology matching—key to knowledge integration—surpassing traditional systems in complex scenarios. Methodologies like Guide-Align mitigate risks associated with LLMs, such as biased content generation, through a two-stage approach employing safety-trained models to ensure contextually appropriate and value-aligned responses [86, 71].

Multi-modal inputs enhance external knowledge integration, enabling LLMs to process diverse information sources, including text, images, and audio, resulting in more comprehensive and contextually appropriate responses [6]. This integration enriches LLM reasoning capabilities and alignment with human values. Verification processes benefit from external knowledge integration, providing

robust frameworks for evaluating LLM output accuracy. Techniques like Bayesian reward models, which incorporate uncertainty estimates, facilitate nuanced assessments of LLM responses, ensuring alignment with human preferences and ethical standards [80].

Moreover, integrating external knowledge supports automated feedback strategies, refining LLM outputs and enhancing alignment with desired outcomes [81]. By systematically incorporating external datasets and domain-specific information, these strategies ensure LLM adaptability to evolving user needs. The integration of external knowledge is vital for advancing AI alignment and verification, equipping LLMs with resources to produce accurate, reliable, and ethically aligned outputs. This approach addresses challenges like hallucination and verifiable accountability while enhancing transparency and reducing biases. By converting domain-specific insights into quantifiable features, LLMs significantly improve predictive analytics and decision-making accuracy, fostering responsible development and prioritizing ethical considerations in AI technologies [12, 24].

5.4 Challenges and Ethical Considerations

The deployment of Large Language Models (LLMs) across various domains, such as healthcare and education, raises significant ethical challenges that require careful management. A primary concern is bias in training data, which can lead to skewed outputs misaligned with human values, perpetuating stereotypes and inaccuracies, especially in sensitive areas like moral decision-making [22]. Ethical considerations extend to transparency, fairness, and user privacy, particularly in conversational recommender systems (CRSs). Ensuring equitable and transparent operations is vital for maintaining user trust and aligning AI outputs with societal norms [41]. The automation of expert knowledge through LLMs also raises ethical questions regarding biases in translating qualitative insights into quantifiable data, necessitating rigorous evaluation and mitigation strategies [12].

Challenges in AI alignment are compounded by limitations in current self-verification methods, which often fail to accurately identify logical fallacies. This underscores the need for robust verification frameworks to address moral alignment nuances and biases from training data [16]. Ethical deployment of LLMs demands comprehensive approaches to mitigate risks associated with user-generated content, which can introduce malicious intent or amplify biases. Addressing challenges like hallucination and accountability, alongside developing tailored ethical frameworks and dynamic auditing systems, enhances transparency and reduces bias. Research indicates that LLMs can inadvertently support misconceptions and provide unreliable information, emphasizing the importance of critical evaluation and responsible interaction with these technologies to ensure safe and equitable use [53, 22, 57, 12, 24]. Future research should focus on resilient alignment processes that mitigate these risks, ensuring LLMs operate within ethical boundaries. Prioritizing these considerations will lead to more aligned and responsible LLMs that adhere to societal and ethical standards.

6 Challenges and Future Directions

6.1 Advancements in Process Supervision Techniques

Recent progress in process supervision techniques has significantly enhanced the performance and adaptability of Large Language Models (LLMs). Frameworks like LLM2LLM highlight inter-LLM knowledge transfer, achieving up to 24.2% improvement in low-data scenarios, as demonstrated on datasets like GSM8K [18]. These advancements underscore the potential of structured problem-solving and diverse solution exploration in data-scarce environments.

Enhancements in keyword extraction capabilities through LLMs contribute to improved contextual understanding and semantic accuracy, crucial for refining process supervision techniques [1]. The integration of multi-modal inputs in LLM-based recommendation systems suggests promising research directions to optimize the balance between model complexity and performance [20].

Challenges persist in calibrating LLM confidence scores, necessitating future research to enhance these techniques across diverse languages and datasets for more reliable outputs [10]. Current reinforcement learning methods often overlook critical output components that determine correctness. Techniques like Reinforcement Learning with Minimum Editing Constraint (RLMEC) provide targeted supervision, emphasizing vital output elements [21].

Exploring the philosophical implications of LLM credences is essential, as advancements in assessing these credences enhance interpretability and trustworthiness [19]. Implementing domain-specific training, iterative feedback mechanisms, and task-specific optimizations will significantly enhance LLM capabilities, addressing modern applications' complex requirements, such as improving table-to-text generation and optimizing performance in low-data environments through innovative data augmentation techniques [18, 29].

6.2 Technological Advancements and Future Research Directions

The evolution of Large Language Models (LLMs) is propelled by technological advancements aimed at enhancing their performance across diverse domains. Future research should focus on refining model architectures and training data structures, particularly in domain-specific applications like medical and recommendation systems. In Conversational Recommender Systems (CRS), improvements in LLM adaptation and expert model integration are crucial for performance enhancement [41].

Enhancing the robustness of routing mechanisms in modular systems such as MRKL is vital, focusing on expanding task ranges and integrating additional knowledge sources [43]. Developing benchmarks with diverse and complex natural language inputs will provide comprehensive evaluations of LLM capabilities [87].

Empirical testing and enhancing encoding techniques are essential for integrating expert intuition into LLMs [12]. Refining frameworks for efficiency, exploring alternative reward models, and applying these approaches to a broader array of reasoning tasks are vital for advancing LLM research [14]. Adapting models like LARM for real-world embodied AI tasks will broaden LLM applicability beyond simulation [6].

In prompt engineering, enhancing techniques and integrating additional LLMs for improved keyword extraction is essential [1]. Addressing prompt hacking challenges requires comprehensive studies across various LLMs to bridge existing knowledge gaps [15]. Advancements in privacy protection capabilities necessitate comprehensive datasets and evaluation methods aligned with real-world privacy challenges [17].

Exploring hyperparameter tuning for frameworks like LLM2LLM, along with integrating techniques such as prompt tuning and few-shot learning, represents promising research avenues [18]. Applying Reinforcement Learning with Minimum Editing Constraint (RLMEC) on larger LLMs to enhance human alignment and reduce hallucinations also requires further investigation [21].

Technological advancements in LLMs and future research directions are crucial for fostering responsible applications. This includes addressing unique ethical challenges such as hallucination, accountability, and bias while enhancing transparency through human-centered approaches. A robust framework for LLM development must incorporate precise specifications to ensure modularity and reliability, guiding integration into various applications while prioritizing ethical considerations and stakeholder needs [23, 62, 24, 29].

6.3 Expanding LLM Applications and Domains

The potential for expanding Large Language Models (LLMs) across various domains is substantial, driven by advancements in integrating diverse knowledge sources and enhancing learning strategies. Significant gaps remain in understanding the long-term impacts of these strategies, particularly regarding the integration of diverse knowledge sources, essential for evolving LLM capabilities [88]. Incorporating Knowledge Graphs is promising for expanding LLM applications, especially in educational contexts, facilitating effective information retrieval and knowledge dissemination [59].

Exploring self-assessment mechanisms within LLMs offers opportunities for enhancing AI behavior and trustworthiness. Future research should focus on a broader set of variables to deepen understanding of how LLMs assess confidence and competence, significantly impacting AI behavior and decision-making [89]. Techniques like SIPF on larger models and effectiveness in a wider range of reasoning tasks beyond mathematics and code generation warrant further investigation, presenting potential for LLMs to address complex, multi-domain challenges [46].

The societal impact of LLMs is profound, with potential applications spanning multilingual tasks requiring robust training strategies and enhanced model architectures for effective continuous learning

from new multilingual data [90]. However, this expansion raises concerns regarding job displacement and misinformation, emphasizing the necessity for responsible AI use and developing interdisciplinary methodologies to evaluate trust in AI systems.

7 Conclusion

This survey highlights the significant impact of Large Language Models (LLMs) across diverse sectors, emphasizing the critical roles of process supervision and AI alignment in enhancing their efficacy and dependability. Advances in reasoning verification, notably through Process Reward Models (PRMs), have markedly improved the success rates of code generation models, particularly in complex, extended tasks. These developments have propelled LLM applications forward, with notable success in the medical domain, where tailored LLMs have shown proficiency comparable to that of medical professionals in clinical text summarization.

The research further reveals that LLMs exhibit varied ethical profiles, with proprietary models often aligning with utilitarian principles, while open models tend to embody values-based ethics. This variation highlights the importance of AI alignment to ensure LLMs operate within ethical frameworks and societal norms. The establishment of a structured framework for evaluating trust in LLM-based systems is crucial, focusing on reliability, transparency, and user interaction to build confidence.

Despite these advancements, challenges remain, particularly in automating systematic reviews, where current limitations necessitate careful application. Additionally, although LLMs perform consistently in competitive markets, they lack the adaptive capabilities of human traders, indicating a need for more advanced AI models.

The strides made in process supervision and AI alignment are pivotal in advancing natural language processing and AI, laying the groundwork for more robust, dependable, and ethically aligned AI systems. These innovations are essential for meeting the complex requirements of modern applications, ensuring that LLMs not only achieve technical superiority but also conform to ethical and societal expectations.

References

[1] Sandeep Chataut, Tuyen Do, Bichar Dip Shrestha Gurung, Shiva Aryal, Anup Khanal, Carol Lushbough, and Etienne Gnimpieba. Comparative study of domain driven terms extraction using large language models, 2024.

[2] Alejandro Peña, Aythami Morales, Julian Fierrez, Ignacio Serna, Javier Ortega-Garcia, Iñigo Puente, Jorge Cordova, and Gonzalo Cordova. Leveraging large language models for topic classification in the domain of public affairs, 2023.

[3] Yuan Tian, Weiwei Cui, Dazhen Deng, Xinjing Yi, Yurun Yang, Haidong Zhang, and Yingcai Wu. Chartgpt: Leveraging llms to generate charts from abstract natural language, 2023.

[4] Xuelong Geng, Tianyi Xu, Kun Wei, Bingshen Mu, Hongfei Xue, He Wang, Yangze Li, Pengcheng Guo, Yuhang Dai, Longhao Li, Mingchen Shao, and Lei Xie. Unveiling the potential of llm-based asr on chinese open-source datasets, 2024.

[5] Sergio Rojas-Galeano. Zero-shot spam email classification using pre-trained large language models, 2024.

[6] Zhuoling Li, Xiaogang Xu, Zhenhua Xu, SerNam Lim, and Hengshuang Zhao. Larm: Large auto-regressive model for long-horizon embodied intelligence, 2025.

[7] Ryan Burnell, Han Hao, Andrew R. A. Conway, and Jose Hernandez Orallo. Revealing the structure of language model capabilities, 2023.

[8] Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. Tool learning with large language models: A survey, 2024.

[9] Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. A survey on human preference learning for large language models, 2024.

[10] Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel J. Candès, and Dan Jurafsky. Can unconfident llm annotations be used for confident conclusions?, 2025.

[11] Wei Wang and Qing Li. Schrodinger's memory: Large language models, 2024.

[12] Phoebe Jing, Yijing Gao, Yuanhang Zhang, and Xianlong Zeng. Translating expert intuition into quantifiable features: Encode investigator domain knowledge via llm for enhanced predictive analytics, 2024.

[13] Ziwei Chai, Tianjie Zhang, Liang Wu, Kaiqiao Han, Xiaohai Hu, Xuanwen Huang, and Yang Yang. Graphllm: Boosting graph reasoning ability of large language model, 2023.

[14] Enhancing llm reasoning with rew.

[15] Baha Rababah, Shang, Wu, Matthew Kwiatkowski, Carson Leung, and Cuneyt Gurcan Akcora. Sok: Prompt hacking of large language models, 2024.

[16] Ruixin Hong, Hongming Zhang, Xinyu Pang, Dong Yu, and Changshui Zhang. A closer look at the self-verification abilities of large language models in logical reasoning, 2024.

[17] Yuqi Yang, Xiaowen Huang, and Jitao Sang. Exploring the privacy protection capabilities of chinese large language models, 2024.

[18] Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala Anumanchipalli, Michael W. Mahoney, Kurt Keutzer, and Amir Gholami. Llm2llm: Boosting llms with novel iterative data enhancement, 2024.

[19] Geoff Keeling and Winnie Street. On the attribution of confidence to large language models, 2024.

[20] Jiahao Tian, Jinman Zhao, Zhenkai Wang, and Zhicheng Ding. Mmrec: Llm based multi-modal recommender system, 2024.

[21] Zhipeng Chen, Kun Zhou, Wayne Xin Zhao, Junchen Wan, Fuzheng Zhang, Di Zhang, and Ji-Rong Wen. Improving large language models via fine-grained reinforcement learning with minimum editing constraint. *arXiv preprint arXiv:2401.06081*, 2024.

[22] Alejandro Tlaie. Exploring and steering the moral compass of large language models, 2024.

[23] Q. Vera Liao and Jennifer Wortman Vaughan. Ai transparency in the age of llms: A human-centered research roadmap, 2023.

[24] Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. Navigating llm ethics: Advancements, challenges, and future directions, 2024.

[25] Jian-Qiao Zhu, Haijiang Yan, and Thomas L. Griffiths. Language models trained to do arithmetic predict human risky and intertemporal choice, 2024.

[26] Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, and Leslie A. Lenert. The emergence of large language models (llm) as a tool in literature reviews: an llm automated systematic review, 2024.

[27] Yang Yan, Yihao Wang, Chi Zhang, Wenyuan Hou, Kang Pan, Xingkai Ren, Zelun Wu, Zhixin Zhai, Enyun Yu, Wenwu Ou, and Yang Song. Llm4pr: Improving post-ranking in search engine with large language models, 2024.

[28] Ankit Satpute, Noah Giessing, Andre Greiner-Petter, Moritz Schubotz, Olaf Teschke, Akiko Aizawa, and Bela Gipp. Can llms master math? investigating large language models on math stack exchange, 2024.

[29] Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. Investigating table-to-text generation capabilities of llms in real-world information seeking scenarios, 2023.

[30] Ning Dai, Zheng Wu, Renjie Zheng, Ziyun Wei, Wenlei Shi, Xing Jin, Guanlin Liu, Chen Dun, Liang Huang, and Lin Yan. Process supervision-guided policy optimization for code generation, 2025.

[31] C. Nicolò De Sabbata, Theodore R. Sumers, and Thomas L. Griffiths. Rational metareasoning for large language models, 2024.

[32] Weimin Xiong, Yifan Song, Xiutian Zhao, Wenhao Wu, Xun Wang, Ke Wang, Cheng Li, Wei Peng, and Sujian Li. Watch every step! llm agent learning via iterative step-level process refinement, 2024.

[33] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. Adapted large language models can outperform medical experts in clinical text summarization, 2024.

[34] Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for llm problem-solving. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*, 2024.

[35] Pei-Fu Guo, Ying-Hsuan Chen, Yun-Da Tsai, and Shou-De Lin. Towards optimizing with large language models, 2024.

[36] Nicholas Ichien, Dušan Stamenković, and Keith J. Holyoak. Large language model displays emergent ability to interpret novel literary metaphors, 2024.

[37] Jingru Jia and Zehua Yuan. An experimental study of competitive market behavior through llms, 2024.

[38] Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models, 2024.

[39] Zhenwen Liang, Ye Liu, Tong Niu, Xiangliang Zhang, Yingbo Zhou, and Semih Yavuz. Improving llm reasoning through scaling inference computation with collaborative verification. *arXiv preprint arXiv:2410.05318*, 2024.

[40] V. K. Cody Bumgardner, Aaron Mullen, Sam Armstrong, Caylin Hickey, and Jeff Talbert. Local large language models for complex structured medical tasks, 2023.

[41] Yue Feng, Shuchang Liu, Zhenghai Xue, Qingpeng Cai, Lantao Hu, Peng Jiang, Kun Gai, and Fei Sun. A large language model enhanced conversational recommender system, 2023.

[42] Weichuan Wang, Zhaoyi Li, Defu Lian, Chen Ma, Linqi Song, and Ying Wei. Mitigating the language mismatch and repetition issues in llm-based machine translation via model editing, 2024.

[43] Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, Dor Muhlgay, Noam Rozen, Erez Schwartz, Gal Shachaf, Shai Shalev-Shwartz, Amnon Shashua, and Moshe Tenenholtz. Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning, 2022.

[44] Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. The impact of reasoning step length on large language models, 2024.

[45] Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models, 2023.

[46] Kaiyuan Chen, Jin Wang, and Xuejie Zhang. Learning to reason via self-iterative process feedback for small language models. *arXiv preprint arXiv:2412.08393*, 2024.

[47] Jianing Qi, Hao Tang, and Zhigang Zhu. Verifierq: Enhancing llm test time compute with q-learning-based verifiers. *arXiv preprint arXiv:2410.08048*, 2024.

[48] Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. Improve mathematical reasoning in language models by automated process supervision, 2024.

[49] S olving m ath w ord p roblems w.

[50] Randall Balestriero, Romain Cosentino, and Sarath Shekkizhar. Characterizing large language model geometry helps solve toxicity detection and generation, 2024.

[51] Menna Fateen and Tsunenori Mine. Developing a tutoring dialog dataset to optimize llms for educational use, 2024.

[52] Dimitrios Michael Manias, Ali Chouman, and Abdallah Shami. Semantic routing for enhanced performance of llm-assisted intent-based 5g core network management and orchestration, 2024.

[53] Milena Chadimová, Eduard Jurášek, and Tomáš Kliegr. Meaningless is better: hashing bias-inducing words in llm prompts improves performance in logical reasoning and statistical learning, 2024.

[54] Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. Data augmentation using large language models: Data perspectives, learning paradigms and challenges, 2024.

[55] Ramya Keerthy Thatikonda, Jiuzhou Han, Wray Buntine, and Ehsan Shareghi. Strategies for improving nl-to-fol translation with llms: Data generation, incremental fine-tuning, and verification, 2024.

[56] Xianlong Zeng, Yijing Gao, Fanghao Song, and Ang Liu. Similar data points identification with llm: A human-in-the-loop strategy using summarization and hidden state insights, 2024.

[57] Yufan Chen, Arjun Arunasalam, and Z. Berkay Celik. Can large language models provide security privacy advice? measuring the ability of llms to refute misconceptions, 2023.

[58] Kevin Wu, Eric Wu, and James Zou. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence, 2025.

[59] Tuan Bui, Oanh Tran, Phuong Nguyen, Bao Ho, Long Nguyen, Thang Bui, and Tho Quan. Cross-data knowledge graph construction for llm-enabled educational question-answering system: A case study at hcmut, 2024.

[60] Zihan Wang, Yunxuan Li, Yuexin Wu, Liangchen Luo, Le Hou, Hongkun Yu, and Jingbo Shang. Multi-step problem solving through a verifier: An empirical analysis on model-induced process supervision, 2024.

[61] Sivan Schwartz, Avi Yaeli, and Segev Shlomov. Enhancing trust in llm-based ai automation agents: New considerations and future challenges, 2023.

[62] Ion Stoica, Matei Zaharia, Joseph Gonzalez, Ken Goldberg, Koushik Sen, Hao Zhang, Anastasios Angelopoulos, Shishir G. Patil, Lingjiao Chen, Wei-Lin Chiang, and Jared Q. Davis. Specifications: The missing link to making the development of llm systems an engineering discipline, 2024.

[63] Fabien Roger and Ryan Greenblatt. Preventing language models from hiding their reasoning, 2023.

[64] Shengyu Feng, Xiang Kong, Shuang Ma, Aonan Zhang, Dong Yin, Chong Wang, Ruoming Pang, and Yiming Yang. Step-by-step reasoning for math problems via twisted sequential monte carlo, 2024.

[65] Melissa Ailem, Katerina Marazopoulou, Charlotte Siska, and James Bono. Examining the robustness of llm evaluation to the distributional assumptions of benchmarks, 2024.

[66] Ziyang Chen and Stylios Moscholios. Using prompts to guide large language models in imitating a real person's language style, 2024.

[67] Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. Llms4ol: Large language models for ontology learning, 2023.

[68] Gloria Ashiya Katuka, Alexander Gain, and Yen-Yun Yu. Investigating automatic scoring and feedback using large language models, 2024.

[69] Qusai Khraisha, Sophie Put, Johanna Kappenberg, Azza Warraitch, and Kristin Hadfield. Can large language models replace humans in the systematic review process? evaluating gpt-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages, 2023.

[70] Aparna Elangovan, Ling Liu, Lei Xu, Sravan Bodapati, and Dan Roth. Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models, 2024.

[71] Yi Luo, Zhenghao Lin, Yuhao Zhang, Jiashuo Sun, Chen Lin, Chengjin Xu, Xiangdong Su, Yelong Shen, Jian Guo, and Yeyun Gong. Ensuring safe and high-quality outputs: A guideline library approach for language models, 2024.

[72] Swapnaja Achintalwar, Adriana Alvarado Garcia, Ateret Anaby-Tavor, Ioana Baldini, Sara E. Berger, Bishwaranjan Bhattacharjee, Djallel Bouneffouf, Subhajit Chaudhury, Pin-Yu Chen, Lamogha Chiazor, Elizabeth M. Daly, Kirushikesh DB, Rogério Abreu de Paula, Pierre Dognin, Eitan Farchi, Soumya Ghosh, Michael Hind, Raya Horesh, George Kour, Ja Young Lee, Nishtha Madaan, Sameep Mehta, Erik Miehling, Keerthiram Murugesan, Manish Nagireddy, Inkit Padhi, David Piorkowski, Ambrish Rawat, Orna Raz, Prasanna Sattigeri, Hendrik Strobelt, Sarathkrishna Swaminathan, Christoph Tillmann, Aashka Trivedi, Kush R. Varshney, Dennis Wei, Shalisha Witherspooon, and Marcel Zalmanovici. Detectors for safe and reliable llms: Implementations, uses, and limitations, 2024.

[73] Shabnam Hassani. Enhancing legal compliance and regulation analysis with large language models, 2024.

[74] Bumjin Park and Jaesik Choi. Memorizing documents with guidance in large language models, 2024.

[75] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*, 2023.

[76] Zihao Wang, Chirag Nagpal, Jonathan Berant, Jacob Eisenstein, Alex D'Amour, Sanmi Koyejo, and Victor Veitch. Transforming and combining rewards for aligning large language models, 2024.

[77] Rana Muhammad Shahroz Khan, Pingzhi Li, Sukwon Yun, Zhenyu Wang, Shahriar Nirjon, Chau-Wai Wong, and Tianlong Chen. Portllm: Personalizing evolving large language models with training-free and portable model patches, 2024.

[78] Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms, 2024.

[79] Huanqian Wang, Yang Yue, Rui Lu, Jingxin Shi, Andrew Zhao, Shenzhi Wang, Shiji Song, and Gao Huang. Model surgery: Modulating llm's behavior via simple parameter editing, 2025.

[80] Adam X. Yang, Maxime Robeyns, Thomas Coste, Zhengyan Shi, Jun Wang, Haitham Bou-Ammar, and Laurence Aitchison. Bayesian reward models for llm alignment, 2024.

[81] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies, 2023.

[82] Zelong Li, Shuyuan Xu, Kai Mei, Wenyue Hua, Balaji Rama, Om Raheja, Hao Wang, He Zhu, and Yongfeng Zhang. Autoflow: Automated workflow generation for large language model agents, 2024.

[83] Qingru Zhang, Xiaodong Yu, Chandan Singh, Xiaodong Liu, Liyuan Liu, Jianfeng Gao, Tuo Zhao, Dan Roth, and Hao Cheng. Model tells itself where to attend: Faithfulness meets automatic attention steering, 2024.

[84] Jinxin Liu, Shulin Cao, Jiaxin Shi, Tingjian Zhang, Lunyiu Nie, Linmei Hu, Lei Hou, and Juanzi Li. How proficient are large language models in formal languages? an in-depth insight for knowledge base question answering, 2024.

[85] Tanushree Banerjee, Richard Zhu, Runzhe Yang, and Karthik Narasimhan. Llms are superior feedback providers: Bootstrapping reasoning for lie detection with self-generated feedback, 2024.

[86] Hamed Babaei Giglou, Jennifer D'Souza, Felix Engel, and Sören Auer. Llms4om: Matching ontologies with large language models, 2024.

[87] Shawn Gavin, Tuney Zheng, Jiaheng Liu, Quehry Que, Noah Wang, Jian Yang, Chenchen Zhang, Wenhao Huang, Wenhu Chen, and Ge Zhang. Longins: A challenging long-context instruction-based exam for llms, 2024.

[88] Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. Towards lifelong learning of large language models: A survey, 2024.

[89] Aniket Kumar Singh, Suman Devkota, Bishal Lamichhane, Uttam Dhakal, and Chandra Dhakal. The confidence-competence gap in large language models: A cognitive study, 2023.

[90] Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. A survey on large language models with multilingualism: Recent advances and new frontiers, 2025.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.