# Multi-Agent Framework for Email Generation: From Pre-trained Models to DPO Fine-tuning

Waris Ratthapoom

*Supervisor:* Dr. Cass Zhixue Zhao

A report submitted in partial fulfilment of the requirements
for the degree of MSc Artificial Intelligence in Computer Science

*in the*

Department of Computer Science

August 8, 2025

## Declaration

All sentences or passages quoted in this document from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure.

Name: ______________________________

Signature: ______________________________

Date: ______________________________

# Abstract

Multi-agent artificial intelligence systems have emerged as a promising paradigm for complex natural language generation tasks, enabling collaborative problem-solving through specialized agent architectures Yan et al. (2025), Guo et al. (2024). Automated email generation faces significant challenges in achieving human-like quality while maintaining consistency across different optimization strategies Zhang & Tetreault (2019), Chen et al. (2019). Direct Preference Optimization (DPO) shows promise for aligning language models with human preferences, yet its effectiveness within multi-agent frameworks remains unexplored Rafailov et al. (2023), Muldrew et al. (2024). Existing evaluation methodologies suffer from significant limitations in objectivity and consistency when comparing multiple model variants, particularly lacking standardized assessment frameworks for multi-agent systems Ye et al. (2024), Gu et al. (2024). Here we show that a three-agent architecture comprising Email Generator, Checklist Creator, and Judge Agent achieves statistically equivalent performance across three DPO variants—Baseline, DPO-Synthetic, and DPO-Hybrid—when evaluated on 50 unseen validation topics ($F(2, 747) = 0.329$, $p = 0.720$, $\eta^2 = 0.001$), demonstrating robustness in maintaining consistent email generation quality regardless of optimization strategy. This statistical equivalence reveals fundamental limitations in DPO effectiveness when constrained by small training datasets (400-425 preference pairs), suggesting dataset size may be more critical than optimization sophistication. The novel Hybrid prompting strategy with reasoning models successfully overcame traditional evaluation biases, establishing a replicable framework for objective multi-agent system assessment. These findings challenge assumptions about the necessity of complex DPO variants and provide evidence that robust multi-agent architectures maintain consistent performance regardless of underlying optimization complexity. For automated content generation, these results suggest architectural robustness may be more crucial than sophisticated optimization techniques when working with constrained datasets, reshaping resource prioritization between system design and data optimization. These findings mark a significant evolution in automated content generation research, demonstrating that sophisticated multi-agent architectures achieve consistent performance independent of optimization complexity, shifting focus from algorithmic sophistication to architectural robustness Ferrag et al. (2025*a*), Liu et al. (2025). The established methodology represents a transferable framework applicable to diverse collaborative AI assessment challenges while providing guidance for future multi-agent system development Masterman et al. (2024), Sapkota et al. (2025).

# Contents

**1 Introduction** **1**
1.1 Multi-Agent Systems in Artificial Intelligence . . . . . . . . . . . . . . . . . . . 1
1.2 Email Generation and Preference Alignment . . . . . . . . . . . . . . . . . . . . 1
1.3 Evaluation Challenges in Multi-Agent Systems . . . . . . . . . . . . . . . . . . 2
1.4 Research Gaps and Methodological Limitations . . . . . . . . . . . . . . . . . . 3
1.5 Research Approach and Methodology Innovation . . . . . . . . . . . . . . . . . 4
1.5.1 Methodology Overview and Technical Architecture . . . . . . . . . . . 4
1.5.2 Technical Innovation: Hybrid Prompting Strategy Development . . . . 5
1.5.3 Model Selection Rationale and Experimental Design . . . . . . . . . . 6
1.6 Research Implications . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7
1.6.1 Model Selection and Resource Allocation Practices . . . . . . . . . . . 7
1.6.2 DPO Application Considerations and Training Data Constraints . . . 7
1.6.3 Evaluation Framework Broader Applicability . . . . . . . . . . . . . . 8
1.7 Research Contribution and Dissertation Overview . . . . . . . . . . . . . . . . . 8
1.7.1 Dissertation Structure and Chapter Organization . . . . . . . . . . . . 9

**2 Literature Review** **11**
2.1 Multi-Agent Evaluation Framework Gaps . . . . . . . . . . . . . . . . . . . . . 11
2.2 DPO Performance Under Data Constraints . . . . . . . . . . . . . . . . . . . . 12
2.3 Bias in AI System Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14
2.4 Email Generation Assessment Challenges . . . . . . . . . . . . . . . . . . . . . 16
2.5 Research Positioning and Contributions . . . . . . . . . . . . . . . . . . . . . . 18

**3 Methodology** **21**
3.1 Research Design and Multi-Agent Architecture . . . . . . . . . . . . . . . . . . 21
3.1.1 Multi-Agent Evaluation Framework . . . . . . . . . . . . . . . . . . . . 21
3.1.2 Three-Agent Architecture Design . . . . . . . . . . . . . . . . . . . . . 21
3.1.3 Agent Interaction and Data Flow . . . . . . . . . . . . . . . . . . . . . 22
3.1.4 Reasoning Model Selection Rationale . . . . . . . . . . . . . . . . . . . 23
3.2 Dataset Development and Model Selection . . . . . . . . . . . . . . . . . . . . . 23
3.2.1 Human Email Foundation and AI-Generated Expansion . . . . . . . . 23
3.2.2 Validation Dataset Creation for Final Assessment . . . . . . . . . . . . 24
3.2.3 Topic Categories and Distribution . . . . . . . . . . . . . . . . . . . . 24
3.2.4 Email Generation Model Selection and Categorization . . . . . . . . . 25
3.2.5 Agent Model Experimentation and Selection . . . . . . . . . . . . . . 25
3.3 Evaluation Framework Development . . . . . . . . . . . . . . . . . . . . . . . . 26
3.3.1 Checklist Agent Prompting Strategy Optimization . . . . . . . . . . . 27
3.3.2 Hybrid Prompting Strategy Validation . . . . . . . . . . . . . . . . . . 27
3.3.3 Hybrid Framework Implementation and Validation . . . . . . . . . . . 27
3.3.4 Binary Checklist Scoring and Judge Agent Integration . . . . . . . . . 29
3.3.5 Framework Reliability and Quality Assurance Integration . . . . . . . 29
3.4 Direct Preference Optimization Implementation . . . . . . . . . . . . . . . . . . 29
3.4.1 Dual-Method DPO Approach Development . . . . . . . . . . . . . . . 30

3.4.2 Preference Pair Generation and Data Preparation . . . . . . . . . . . 30
3.4.3 DPO-Synthetic Method Implementation . . . . . . . . . . . . . . . . . . . 30
3.4.4 DPO-Hybrid Method Implementation . . . . . . . . . . . . . . . . . . . . 31
3.5 Final Validation Protocol . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31
3.5.1 Three-Way Comparison Experimental Design . . . . . . . . . . . . . . 31
3.5.2 Unseen Topic Evaluation Methodology . . . . . . . . . . . . . . . . . . 32
3.5.3 Statistical Analysis Framework . . . . . . . . . . . . . . . . . . . . . . . 32
3.5.4 External Validation and Expert Assessment . . . . . . . . . . . . . . . 32
3.5.5 Generalizability Assessment and Limitations . . . . . . . . . . . . . . 33

**4 Results 36**
4.1 Empirical Overview . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36
4.2 Descriptive Statistics . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36
4.3 Inferential Statistical Analysis . . . . . . . . . . . . . . . . . . . . . . . . . 37
4.4 Effect Size Quantification . . . . . . . . . . . . . . . . . . . . . . . . . . . 38
4.5 Model-Specific Performance Patterns . . . . . . . . . . . . . . . . . . . . . 39
4.6 Domain-Specific Performance Analysis . . . . . . . . . . . . . . . . . . . . 41
4.7 Predictive Validity Assessment . . . . . . . . . . . . . . . . . . . . . . . . 42

**5 Discussion 43**
5.1 Statistical Equivalence Analysis . . . . . . . . . . . . . . . . . . . . . . . . 43
5.2 Framework Robustness Evaluation . . . . . . . . . . . . . . . . . . . . . . 44
5.3 Theoretical Implications and Methodological Contributions . . . . . . . . . . 46

**6 Conclusion 48**
6.1 Summary of Key Findings . . . . . . . . . . . . . . . . . . . . . . . . . . . 48
6.1.1 Primary Research Outcomes . . . . . . . . . . . . . . . . . . . . . . . 48
6.1.2 Model-Specific Response Heterogeneity . . . . . . . . . . . . . . . . . 48
6.1.3 Framework Robustness Validation . . . . . . . . . . . . . . . . . . . . 48
6.1.4 Methodological Innovation Outcomes . . . . . . . . . . . . . . . . . . . 48
6.2 Future Research Directions . . . . . . . . . . . . . . . . . . . . . . . . . . 49
6.2.1 Short-Term Research Extensions . . . . . . . . . . . . . . . . . . . . . 49
6.2.2 Medium-Term Research Agenda . . . . . . . . . . . . . . . . . . . . . 49
6.2.3 Technical and Methodological Advances . . . . . . . . . . . . . . . . . 50
6.2.4 Long-Term Research Vision . . . . . . . . . . . . . . . . . . . . . . . . 50
6.3 Final Remarks . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 51

**Appendices 58**

**A Experimental Setup Details 59**

# List of Figures

3.1 Multi-Agent System Architecture . . . . . . . . . . . . . . . . . . . . . . . . . . . 22
3.2 Agent Model Selection Comparison . . . . . . . . . . . . . . . . . . . . . . . . . . 26
3.3 Hybrid Evaluation Framework Workflow . . . . . . . . . . . . . . . . . . . . . . 28
3.4 Final Validation Protocol . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34

4.1 Model Performance Comparison . . . . . . . . . . . . . . . . . . . . . . . . . . . 37
4.2 ANOVA Results Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 38
4.3 Effect Size Forest Plot with 95% Confidence Intervals . . . . . . . . . . . . . . 39
4.4 Model-Specific Improvement Forest Plot . . . . . . . . . . . . . . . . . . . . . . 40
4.5 Model Size Group Performance Comparison . . . . . . . . . . . . . . . . . . . . 41

# List of Tables

3.1 Topic Dataset Distribution . . . . . 25
3.2 DPO Preference Pair Generation Summary . . . . . 31
3.3 Final Validation Statistical Framework . . . . . 35

4.1 Pairwise Statistical Comparisons Between Model Variants . . . . . 37
4.2 Individual Model Performance by Variant . . . . . 39
4.3 Performance by Topic Category . . . . . 41
4.4 Methodology Validation: Predicted vs Actual Results . . . . . 42

# Chapter 1

# Introduction

## 1.1 Multi-Agent Systems in Artificial Intelligence

The evolution of artificial intelligence systems has increasingly shifted toward collaborative architectures where multiple specialized agents work together to accomplish complex tasks that exceed the capabilities of individual models Guo et al. (2024), Yan et al. (2025). This paradigmatic shift reflects a growing recognition that the challenges facing modern AI applications—particularly in natural language processing—often require diverse expertise, varied perspectives, and sophisticated coordination mechanisms that are difficult to achieve through monolithic approaches Talebirad & Nadiri (2023), Krishnan (2025*b*). Multi-agent systems offer unique advantages in terms of modularity, specialization, and robustness, enabling the development of AI frameworks that can adapt to complex, domain-specific requirements while maintaining interpretability and controllability Ma et al. (2024), Cemri et al. (2025).

Natural language generation has emerged as one of the most promising applications for multi-agent architectures, particularly in domains requiring high levels of human-like communication and contextual understanding Pauli et al. (2024), Murakami et al. (2023). The complexity of generating coherent, contextually appropriate, and stylistically consistent text across different domains and purposes has driven researchers to explore collaborative approaches where specialized agents handle distinct aspects of the generation process. Email generation represents a particularly challenging domain within natural language generation due to its requirement for personalized tone, appropriate formality levels, contextual relevance, and persuasive effectiveness—qualities that demand sophisticated understanding of both linguistic conventions and human communication preferences Zhang & Tetreault (2019), Chen et al. (2019).

## 1.2 Email Generation and Preference Alignment

The challenges inherent in automated email generation extend beyond traditional text generation metrics to encompass aspects of pragmatic effectiveness, cultural sensitivity, and recipient-specific adaptation. Traditional approaches to email automation, exemplified by systems like Gmail's Smart Compose Chen et al. (2019), have focused primarily on efficiency and basic coherence rather than sophisticated evaluation of content quality and recipient appropriateness. However, as organizations increasingly rely on automated communication systems for customer engagement, marketing campaigns, and stakeholder relations, the demand for more nuanced and effective email generation systems has grown substantially Henderson et al. (2017), Li, Sun, Yuan, Fan, Zhao & Liu (2023). This evolution necessitates not only improved generation capabilities but also more sophisticated evaluation methodologies that can assess multiple dimensions of email effectiveness simultaneously.

The alignment of artificial intelligence systems with human preferences has become a central concern in the development of practical AI applications, particularly those involving direct human interaction through natural language Rafailov et al. (2023), Wang et al.

(2024). Traditional approaches to model alignment, such as Reinforcement Learning from Human Feedback (RLHF), have proven effective but suffer from computational complexity, training instability, and difficulties in scaling to diverse preference criteria. Direct Preference Optimization (DPO) has emerged as a promising alternative that addresses many of these limitations by directly optimizing models based on preference data without requiring explicit reward model training Muldrew et al. (2024), Gallego (2024). The theoretical elegance and practical advantages of DPO have led to widespread adoption across various natural language processing tasks, yet its effectiveness within multi-agent frameworks remains largely unexplored.

## 1.3 Evaluation Challenges in Multi-Agent Systems

The evaluation of multi-agent systems presents unique methodological challenges that differ significantly from single-model assessment approaches Yehudai et al. (2025), Zhu et al. (2024). Traditional evaluation metrics for natural language generation, such as BLEU scores or perplexity measures, fail to capture the collaborative dynamics, inter-agent communication effectiveness, and emergent behaviors that characterize multi-agent performance Schmidtová et al. (2024), Liu et al. (2023). Furthermore, the evaluation of email generation systems requires domain-specific metrics that assess not only linguistic quality but also pragmatic effectiveness, persuasive impact, and appropriateness for specific communication contexts Li, Sun, Yuan, Fan, Zhao & Liu (2023), Rony et al. (2022).

Existing evaluation approaches in natural language processing suffer from several fundamental limitations that compromise their effectiveness for multi-model comparison Maharana et al. (2023), Ni et al. (2024). Cross-task inconsistency represents a major challenge, where models demonstrate varying performance patterns across different evaluation scenarios, making it difficult to establish reliable comparative assessments Maharana et al. (2023). Additionally, current evaluation frameworks exhibit significant biases related to position preference, superficial reasoning cues, and inconsistent grading standards that undermine the objectivity required for systematic multi-model comparison Ye et al. (2024), Wang et al. (2025). These methodological shortcomings are particularly problematic when evaluating AI systems intended for real-world deployment, where consistent and reliable performance assessment is crucial for informed decision-making Ni et al. (2024).

The challenge is further compounded by the lack of standardized protocols for assessing multi-agent systems, where traditional single-model evaluation metrics prove inadequate for capturing the complex interactions and emergent behaviors that characterize collaborative AI architectures Li, Jiang, Huang, Beigi, Zhao, Tan, Bhattacharjee & Jiang (2024), Xu et al. (2025). Contextual assessment presents additional complications, as evaluation criteria often depend on practitioner priorities and domain-specific requirements, leading to conditional evaluation frameworks that are difficult to standardize across different applications Xu et al. (2025). The development of robust evaluation frameworks for multi-agent email generation systems thus represents a significant methodological challenge that requires careful consideration of bias mitigation, consistency enhancement, and domain-specific assessment criteria.

Recent advances in reasoning-enhanced evaluation approaches have shown promise for improving the assessment of complex AI systems by incorporating explicit reasoning steps and multi-perspective analysis Marjanović et al. (2025), Sui et al. (2025). These approaches leverage the capacity of advanced language models to provide detailed explanations and justifications for their evaluative judgments, potentially offering more transparent and comprehensive assessment of system performance. The integration of reasoning-enhanced evaluation within multi-agent frameworks represents a natural evolution that could significantly improve the reliability and interpretability of performance assessments while providing valuable insights into system behavior and areas for improvement.

## 1.4 Research Gaps and Methodological Limitations

Despite substantial progress in multi-agent systems, DPO optimization, and email generation independently, the intersection of these three domains remains underexplored, revealing several critical research gaps. First, existing research has not systematically investigated how different DPO training strategies perform within multi-agent email generation frameworks, particularly when constrained by limited training data Feng et al. (2024), Deng et al. (2025). Theoretical analyses suggest that DPO optimization may face fundamental limitations when applied to small datasets, as the method's effectiveness depends critically on the quality and quantity of preference pairs available for training Feng et al. (2024). This constraint is particularly relevant for domain-specific applications like email generation, where obtaining large-scale, high-quality preference data can be prohibitively expensive or logistically challenging.

Second, the absence of standardized evaluation frameworks for multi-agent systems creates significant methodological gaps that hinder systematic comparison of different optimization approaches Li, Jiang, Huang, Beigi, Zhao, Tan, Bhattacharjee & Jiang (2024), Gu et al. (2024). Current evaluation methodologies in the field suffer from inconsistent standards, varying protocols across different research communities, and significant biases that compromise the reliability of comparative assessments Ni et al. (2024), Gao et al. (2023). The lack of objective assessment methodologies is particularly problematic for multi-agent systems, where the complex interactions between specialized agents require sophisticated evaluation frameworks capable of capturing both individual agent performance and collective system behavior.

Third, existing approaches to preference optimization have not adequately addressed the unique challenges posed by multi-agent architectures, where the optimization of individual agents may not translate directly to improved system-level performance. The theoretical understanding of how DPO variants perform within collaborative frameworks remains limited, particularly regarding questions of convergence, stability, and the interaction effects between multiple optimized agents Feng et al. (2024), Karthik et al. (2024). This gap represents both a significant research opportunity and a practical necessity given the increasing deployment of AI systems in communication-intensive applications where reliable performance assessment is crucial.

The systematic evaluation of DPO variants within multi-agent email generation frameworks could provide crucial insights into the effectiveness of different alignment strategies and inform the development of more sophisticated automated communication systems. However, the current state of evaluation methodology presents significant obstacles to conducting such systematic comparisons, necessitating the development of novel assessment frameworks that can reliably differentiate between competing optimization approaches while maintaining objectivity and consistency across different experimental conditions.

## 1.5 Research Approach and Methodology Innovation

This research addresses the identified challenges through a comprehensive three-agent architecture that separates the complex task of email generation and evaluation into specialized, collaborative components. The Email Generator agent focuses exclusively on producing contextually appropriate and persuasive email content, while the Checklist Creator agent develops domain-specific evaluation criteria tailored to each communication scenario. The Judge Agent then applies these criteria systematically to assess email quality across multiple dimensions, creating a structured evaluation pipeline that reduces bias and enhances assessment consistency Li, Jiang, Huang, Beigi, Zhao, Tan, Bhattacharjee & Jiang (2024), Rony et al. (2022).

The DPO optimization strategy implemented in this study systematically compares three distinct approaches to preference alignment: a Baseline variant using standard pre-trained models, a DPO-Synthetic variant trained on artificially generated preference pairs, and a DPO-Hybrid variant combining synthetic and human-curated preference data. This controlled comparison enables precise assessment of how different training data compositions affect model performance within the multi-agent framework, while maintaining consistent experimental conditions across all variants Rafailov et al. (2023), Feng et al. (2024). The constraint of limited training data (400-425 preference pairs per variant) reflects realistic resource limitations faced by practitioners implementing such systems in specialized domains.

The evaluation methodology innovation centers on a novel Hybrid prompting strategy that integrates reasoning-enhanced assessment with traditional scoring mechanisms. This approach leverages advanced reasoning models to provide explicit justifications for evaluative decisions, thereby addressing the transparency and consistency limitations that plague existing evaluation frameworks Marjanović et al. (2025), Xu et al. (2025). By incorporating structured reasoning processes into the evaluation pipeline, this methodology reduces position bias, enhances inter-rater reliability, and provides detailed insights into the factors driving system performance across different optimization conditions.

### 1.5.1 Methodology Overview and Technical Architecture

The three-agent architecture developed in this research implements a sophisticated division of labor that mirrors effective human collaboration in content creation and evaluation processes. The Email Generator agent, implemented using transformer-based language models ranging from 1.1B to 70B parameters, specializes in producing contextually appropriate and persuasive email content based on specific charity fundraising scenarios Zhou et al. (2025), Ke et al. (2025). This agent operates through structured prompting that incorporates recipient demographics, organizational context, and communication objectives to generate coherent, engaging email content that maintains appropriate tone and messaging effectiveness.

The Checklist Creator agent functions as a dynamic evaluation framework generator, developing domain-specific assessment criteria tailored to each unique communication scenario Cheng et al. (2024), Qiao et al. (2022). Rather than applying static evaluation rubrics, this agent analyzes the specific context of each email generation task and constructs comprehensive evaluation checklists that capture relevant quality dimensions including clarity, persuasiveness, appropriateness, and factual accuracy. This approach ensures that evaluation criteria remain contextually relevant while maintaining consistency across different topics and scenarios.

The Judge Agent serves as the systematic evaluator, applying the dynamically generated checklists to assess email quality across multiple dimensions using probability-based scoring mechanisms Hadji-Kyriacou & Arandjelovic (2024), Xu et al. (2023). The reasoning model selection rationale centers on leveraging models specifically trained for analytical thinking and structured evaluation, ensuring that assessment decisions incorporate explicit reasoning steps that enhance transparency and reliability. This agent implementation reduces subjective bias through structured evaluation protocols while providing detailed justifications for scoring decisions that enable systematic analysis of performance patterns.

The validation approach employs 50 carefully selected unseen topics that span diverse charity fundraising scenarios, ensuring comprehensive evaluation across different communication contexts while maintaining experimental rigor Gao et al. (2024), Shao et al. (2023). This validation strategy enables robust assessment of system generalizability while providing sufficient statistical power to detect meaningful performance differences between different optimization approaches. The systematic application of this three-agent architecture across multiple model variants and optimization conditions provides unprecedented insights into multi-agent system behavior under controlled experimental conditions.

### 1.5.2 Technical Innovation: Hybrid Prompting Strategy Development

The development of the Hybrid prompting strategy represents a fundamental advancement in multi-agent system evaluation, addressing critical methodological limitations that have historically compromised the reliability of comparative assessments in collaborative AI systems. Traditional evaluation approaches suffer from position bias, inconsistent scoring standards, and superficial reasoning that undermines objective assessment Wang et al. (2025), Ye et al. (2024). The Hybrid prompting strategy mitigates these limitations through a sophisticated integration of structured reasoning processes with probability-based scoring mechanisms that enhance both transparency and consistency.

The probability-based scoring methodology positions evaluation decisions within a statistical framework that enables systematic comparison across different model variants while maintaining objective assessment standards Li, Sun, Yuan, Fan, Zhao & Liu (2023), Liu et al. (2023). Rather than relying on subjective qualitative assessments, this approach quantifies evaluation outcomes through probabilistic measures that facilitate rigorous statistical analysis. The integration of reasoning models ensures that scoring decisions incorporate explicit analytical steps, providing detailed justifications that enhance the interpretability of evaluation outcomes while enabling systematic identification of performance patterns across different optimization conditions.

The experimental design rigor implemented through this Hybrid prompting approach establishes a new standard for multi-agent system evaluation that addresses the specific challenges inherent in collaborative AI assessment Chan et al. (2023), Chen et al. (2024). By incorporating structured reasoning processes into the evaluation pipeline, this methodology reduces the variability and bias that typically compromise multi-model comparisons. The systematic application of this evaluation framework across multiple DPO variants and diverse validation scenarios demonstrates its effectiveness in maintaining consistent assessment standards while providing detailed insights into system performance characteristics.

The technical innovation extends beyond individual component improvements to encompass a comprehensive evaluation ecosystem that supports reproducible, objective assessment of multi-agent systems under diverse experimental conditions Biderman et al. (2024), Siegel et al. (2024). This framework enables systematic investigation of optimization strategies, architectural choices, and performance trade-offs that would be difficult to assess using conventional evaluation methodologies. The establishment of this evaluation protocol represents a significant methodological contribution that facilitates more rigorous and reliable assessment of collaborative AI systems across diverse application domains.

### 1.5.3 Model Selection Rationale and Experimental Design

The selection of language models spanning 1.1B to 70B parameters reflects a strategic approach to comprehensive evaluation that encompasses the full spectrum of computational resources and performance characteristics available to practitioners in real-world deployment scenarios Zhang et al. (2023), Herel & Mikolov (2023). This range enables systematic investigation of how model scale affects multi-agent system performance, providing crucial insights into the relationship between computational investment and collaborative effectiveness. The inclusion of both small-scale models (1.1B-1.6B parameters) and large-scale models (34B-70B parameters) facilitates analysis of performance scaling patterns and identifies optimal resource allocation strategies for different deployment contexts.

The comprehensive model range addresses a critical gap in existing multi-agent system research, where evaluations typically focus on single model architectures or limited parameter ranges that fail to capture the full spectrum of available options Urlana et al. (2024), Pimentel et al. (2024). By systematically evaluating models across this parameter range, this research provides empirical evidence for performance scaling relationships that inform practical deployment decisions. The diversity of model architectures included—ranging from efficient small-scale models like TinyLlama (1.1B) to sophisticated large-scale models like Llama-3-70B—ensures that findings remain generalizable across different computational constraints and performance requirements.

The DPO variant comparison approach implements a controlled experimental design that isolates the effects of different preference optimization strategies while maintaining consistent architectural and evaluation frameworks Feng et al. (2024), Deng et al. (2025). The three variants—Baseline (standard pre-trained models), DPO-Synthetic (artificially generated preference pairs), and DPO-Hybrid (combining synthetic and human-curated data)—represent distinct approaches to preference alignment that reflect different resource allocation strategies available to practitioners. This systematic comparison enables precise assessment of optimization effectiveness under realistic resource constraints.

The experimental design connects directly to comprehensive evaluation objectives by ensuring that performance assessments capture the full range of system behaviors under different optimization conditions Card et al. (2020), Connolly et al. (2023). The constraint of limited training data (400-425 preference pairs per variant) reflects realistic limitations faced by organizations implementing specialized AI systems, where obtaining large-scale preference data can be prohibitively expensive or logistically challenging. This constraint enables investigation of DPO effectiveness under conditions that mirror real-world deployment scenarios, providing practical insights for system designers working within similar limitations.

The systematic evaluation across multiple model variants and optimization approaches establishes a comprehensive empirical foundation for understanding multi-agent system behavior that extends beyond specific implementation details to encompass general principles of collaborative AI system design and optimization Kim et al. (2025), Siegel et al. (2024). This methodological approach ensures that research findings remain applicable across diverse application domains while providing specific guidance for practitioners implementing similar systems in resource-constrained environments.

## 1.6 Research Implications

The findings of this research carry significant implications for multiple dimensions of AI system development and deployment, fundamentally challenging existing paradigms in model selection, preference optimization, and evaluation methodology. The statistical equivalence demonstrated across DPO variants ($F(2, 747) = 0.329$, $p = 0.720$, $\eta^2 = 0.001$) provides crucial evidence that sophisticated optimization techniques may offer diminishing returns when constrained by limited training data, fundamentally reshaping how practitioners approach resource allocation in multi-agent system development Deng et al. (2025), Feng et al. (2024).

### 1.6.1 Model Selection and Resource Allocation Practices

The empirical evidence presented in this study demonstrates that architectural robustness in multi-agent systems can effectively compensate for optimization sophistication, suggesting that development resources may be more efficiently allocated toward system design and integration rather than complex preference optimization strategies Ferrag et al. (2025*a*), Liu et al. (2025). This finding challenges the prevailing industry focus on increasingly sophisticated training methodologies and suggests that practitioners working with constrained datasets should prioritize collaborative architecture development over optimization complexity. The consistent performance across model scales from 1.1B to 70B parameters further indicates that smaller, more computationally efficient models may achieve comparable results within well-designed multi-agent frameworks, providing significant cost savings for deployment scenarios where computational resources are limited Masterman et al. (2024).

The implications for model selection extend beyond individual system performance to encompass broader considerations of scalability, maintainability, and interpretability that are crucial for real-world deployment Sapkota et al. (2025). Multi-agent architectures that maintain consistent performance across different model variants offer practitioners greater flexibility in adapting to changing computational constraints, regulatory requirements, and performance objectives without requiring complete system redesign. This architectural resilience represents a fundamental advantage that may prove more valuable than marginal performance improvements achieved through sophisticated optimization techniques.

### 1.6.2 DPO Application Considerations and Training Data Constraints

The revealed limitations of DPO effectiveness under constrained data conditions (400-425 preference pairs) provide critical guidance for practitioners considering preference optimization in specialized domains where large-scale data collection is prohibitively expensive or logistically challenging Feng et al. (2024), Karthik et al. (2024). The research demonstrates that below certain data thresholds, the complexity of DPO variants offers no measurable advantage over simpler baseline approaches, suggesting that investment in data quality and quantity may yield superior returns compared to algorithmic sophistication.

These findings have particular relevance for specialized applications in domains such as legal document generation, medical communication, technical writing, and other professional contexts where preference data is inherently scarce and expensive to obtain Bernard et al. (2024). Rather than pursuing complex optimization strategies with limited data, practitioners in these domains may achieve better results by focusing on architectural improvements, better prompt engineering, and systematic evaluation frameworks that can reliably assess performance across diverse scenarios.

The research also highlights the importance of considering the interaction between training data constraints and optimization complexity in the broader context of AI system development lifecycle costs Zeng et al. (2023). The computational overhead associated with complex DPO training may not justify the marginal performance improvements achievable with limited preference data, suggesting that simpler approaches may offer superior cost-effectiveness for many practical applications.

### 1.6.3 Evaluation Framework Broader Applicability

The novel Hybrid prompting evaluation methodology developed in this research addresses fundamental limitations in existing multi-agent system assessment approaches and demonstrates transferability across diverse collaborative AI applications Lee & Hockenmaier (2025), Patil (2025). The framework's success in overcoming traditional evaluation biases while maintaining consistency across different experimental conditions establishes a new standard for objective assessment of complex AI systems that extends well beyond email generation applications.

The methodology's emphasis on reasoning-enhanced evaluation protocols addresses critical gaps in current assessment approaches that fail to capture the collaborative dynamics essential to multi-agent system performance Xu et al. (2025). This framework provides a foundation for systematic evaluation across domains ranging from scientific research assistance to creative content generation, automated customer service, and collaborative decision-making systems. The demonstrated reliability and objectivity of the evaluation approach offers practitioners a standardized methodology for comparing different system configurations and optimization strategies while maintaining transparency and interpretability.

The broader applicability of this evaluation framework has particular significance for advancing the field of collaborative AI systems, where traditional single-model assessment metrics prove inadequate for capturing emergent behaviors and inter-agent coordination effectiveness Yehudai et al. (2025). By providing a replicable template for objective assessment, this methodology enables more rigorous comparative studies across different research groups and application domains, facilitating the systematic accumulation of knowledge about multi-agent system design principles and optimization strategies.

## 1.7 Research Contribution and Dissertation Overview

This dissertation addresses these fundamental challenges by developing and evaluating a comprehensive three-agent framework for email generation that integrates state-of-the-art DPO optimization with novel evaluation methodologies designed to overcome the limitations identified in existing approaches. The primary contribution of this research is the establishment of a novel Hybrid prompting evaluation framework that successfully addresses the objectivity and consistency limitations plaguing existing multi-agent system assessment approaches, demonstrating superior reliability compared to conventional evaluation methodologies Li, Jiang, Huang, Beigi, Zhao, Tan, Bhattacharjee & Jiang (2024), Xu et al. (2025).

The research investigates the comparative effectiveness of three DPO variants—Baseline, DPO-Synthetic, and DPO-Hybrid—within this structured multi-agent architecture, revealing statistically equivalent performance across all optimization strategies ($F(2, 747) = 0.329$, $p = 0.720$, $\eta^2 = 0.001$) when evaluated on 50 diverse validation topics. This unexpected finding challenges conventional assumptions about DPO effectiveness in constrained data scenarios and provides crucial insights into the relationship between training data quantity and optimization sophistication. The methodological framework developed in this study offers a replicable template for evaluating multi-agent systems across diverse natural language generation domains, extending beyond email generation to encompass any application requiring collaborative AI assessment with objective, bias-resistant evaluation protocols. Through systematic evaluation using both traditional metrics and novel reasoning-enhanced approaches, this work establishes empirical foundations for understanding multi-agent system behavior under different optimization conditions and provides practical guidance for practitioners implementing similar systems in resource-constrained environments.

### 1.7.1 Dissertation Structure and Chapter Organization

This dissertation follows a systematic progression from theoretical foundations through empirical investigation to practical implications, designed to provide comprehensive understanding of multi-agent systems for email generation and their evaluation. The structure establishes logical connections between methodological development, experimental design, and analytical findings that collectively address the research objectives outlined in this introduction.

Chapter 2 presents a comprehensive literature review that positions this research within the broader context of multi-agent artificial intelligence, Direct Preference Optimization, and automated email generation systems. This chapter establishes the theoretical foundations necessary for understanding the methodological innovations introduced in subsequent chapters and identifies the specific research gaps that this dissertation addresses. The literature review emphasizes recent advances in collaborative AI architectures and evaluation methodologies, providing essential context for the empirical findings presented later in the dissertation Ferrag et al. (2025*a*), Masterman et al. (2024).

Chapter 3 details the methodology employed in developing and evaluating the three-agent architecture, with particular emphasis on the novel Hybrid prompting evaluation strategy that addresses traditional biases in multi-agent system assessment. This chapter provides comprehensive documentation of the experimental design, including the systematic comparison of DPO variants, model selection rationale, and validation procedures employed to ensure reproducible and reliable results. The methodology chapter establishes the empirical foundation for understanding how architectural decisions and optimization strategies interact within collaborative AI frameworks.

Chapter 4 presents the results of the comprehensive evaluation across three DPO variants and multiple model architectures, documenting the statistical equivalence that challenges conventional assumptions about optimization effectiveness in constrained data scenarios. The results chapter provides detailed analysis of performance patterns across different model scales and optimization conditions, supported by rigorous statistical testing and effect size analysis that enables robust interpretation of the empirical findings. This chapter demonstrates the effectiveness of the proposed evaluation methodology while revealing unexpected insights into the relationship between architectural robustness and optimization sophistication.

Chapter 5 discusses the implications of these findings for multi-agent system design, preference optimization strategies, and evaluation methodologies in artificial intelligence research. The discussion integrates the empirical results with broader theoretical considerations and practical applications, exploring how these findings reshape understanding of resource allocation priorities in collaborative AI development. This chapter also addresses limitations of the current research and identifies specific directions for future investigation that build upon the methodological and empirical contributions established in this dissertation.

The concluding chapter synthesizes the key contributions of this research and establishes its significance within the evolving landscape of multi-agent artificial intelligence systems. The conclusion emphasizes the transferability of the methodological innovations to other domains requiring collaborative AI assessment and highlights the practical implications for practitioners working with constrained training data. This chapter also discusses the broader impact of these findings on the field of automated content generation and provides specific recommendations for future research directions that address the fundamental questions raised by this investigation.

# Chapter 2

# Literature Review

## 2.1 Multi-Agent Evaluation Framework Gaps

Multi-agent systems represent a paradigm shift in artificial intelligence, promising enhanced problem-solving capabilities through collaborative agent interactions Guo et al. (2024). However, the evaluation of these systems presents fundamental challenges that current assessment frameworks inadequately address. The coordination dynamics between multiple autonomous agents introduce evaluation complexities absent in single-agent architectures, necessitating novel approaches to measure system-level performance rather than individual component effectiveness. Traditional evaluation metrics, designed for single-model assessments, fail to capture the emergent behaviours and coordination patterns that define multi-agent system success Yan et al. (2025).

The assessment of coordination effectiveness remains particularly problematic, as existing frameworks cannot adequately measure the quality of inter-agent communication and collaborative decision-making processes. Current evaluation approaches predominantly focus on task completion metrics whilst neglecting the underlying coordination mechanisms that enable multi-agent collaboration Ma et al. (2024). This measurement gap becomes increasingly critical when evaluating systems where agent specialisation and role differentiation are fundamental design principles, as traditional assessment methods cannot distinguish between successful coordination and coincidental task completion.

A comprehensive analysis of multi-agent system failures reveals systematic blindspots in current evaluation methodologies Cemri et al. (2025). Cemri et al. identify fourteen distinct failure modes organised into three categories: specification and system design failures, inter-agent misalignment, and task verification and termination challenges. Their taxonomy demonstrates that existing evaluation frameworks consistently overlook critical failure points, particularly those related to agent coordination protocols and communication effectiveness. The study reveals that conventional assessment approaches cannot detect specification failures that manifest only during agent interactions, highlighting the inadequacy of component-level testing for multi-agent systems.

Coordination protocol evaluation presents another significant gap in current assessment frameworks. Krishnan's examination of Model Context Protocol implementations demonstrates that standardised coordination mechanisms require specialised evaluation approaches that current benchmarks cannot provide Krishnan (2025*a*). The research reveals that coordination effectiveness depends on context management, protocol adherence, and dynamic adaptation capabilities—factors that existing evaluation frameworks cannot systematically assess. This limitation becomes particularly pronounced when evaluating systems with heterogeneous agents, where coordination complexity increases exponentially with agent diversity.

Benchmark limitations further compound these evaluation challenges, as demonstrated by recent comprehensive assessments of multi-agent evaluation systems Zhu et al. (2025). The MultiAgentBench framework reveals that existing benchmarks fail to capture collaboration and competition dynamics essential for understanding multi-agent system performance. Current assessment approaches predominantly employ task-specific metrics that cannot generalise across different coordination scenarios, limiting their utility for comprehensive system evaluation. The research demonstrates that milestone-based performance indicators, whilst more informative than binary task completion measures, still inadequately capture the nuanced coordination patterns that characterise effective multi-agent collaboration.

The measurement inadequacies extend beyond coordination assessment to fundamental questions about statistical equivalence in multi-agent performance evaluation. Existing frameworks lack the statistical sophistication necessary to distinguish between genuine performance differences and random variation in multi-agent system outputs. This limitation becomes critical when comparing different agent architectures or coordination strategies, as traditional significance testing approaches may fail to detect meaningful differences in coordination effectiveness whilst simultaneously overclaiming differences that result from measurement noise rather than genuine system capabilities.

These evaluation framework gaps directly relate to the present study's investigation of a three-agent email generation system, where coordination between Email Generator, Checklist Creator, and Judge Agent represents the core system functionality. The statistical equivalence finding of $F(2, 747) = 0.329$, $p = 0.720$, $\eta^2 = 0.001$ across three system variants demonstrates the practical implications of evaluation framework limitations. This result suggests either genuine performance equivalence between system variants or, alternatively, measurement inadequacy that obscures meaningful differences in coordination effectiveness. The negligible effect size ($\eta^2 = 0.001$) particularly highlights the challenge of detecting coordination quality differences using conventional statistical approaches, illustrating the broader evaluation framework gaps identified in the literature.

Addressing these evaluation challenges requires developing assessment frameworks that can systematically measure coordination effectiveness, protocol adherence, and emergent system behaviours whilst providing statistical approaches capable of distinguishing genuine performance differences from measurement artifacts. The literature demonstrates a clear need for evaluation methodologies specifically designed for multi-agent systems, moving beyond adapted single-agent assessment approaches toward frameworks that recognise the fundamental complexity of multi-agent coordination dynamics.

## 2.2 DPO Performance Under Data Constraints

Direct Preference Optimization has emerged as a transformative approach for aligning language models with human preferences, yet its effectiveness fundamentally depends on the availability of sufficient high-quality preference data Rafailov et al. (2023). The scalability limitations of DPO under constrained datasets represent a critical gap in current literature, with mounting evidence suggesting that performance gains diminish substantially when preference datasets fall below specific threshold sizes. This data scarcity problem becomes particularly acute in specialised domains where collecting large-scale preference annotations proves prohibitively expensive or practically infeasible, leading to questions about the broader applicability of DPO-based alignment strategies.

The mathematical foundations of DPO performance scaling reveal concerning patterns when dataset constraints are considered. Recent theoretical analysis demonstrates that

DPO's gradient vector field exhibits asymmetric behaviour, decreasing the probability of dispreferred responses faster than it increases preferred response probabilities Feng et al. (2024). This asymmetry becomes pronounced under data limitations, as the optimisation process lacks sufficient examples to establish stable preference boundaries. Feng et al. identify that DPO's sensitivity to supervised fine-tuning effectiveness compounds exponentially when preference datasets contain fewer than 500 high-quality pairs, creating a cascade effect where initial training inadequacies are amplified rather than corrected through preference optimisation.

Empirical investigations of data selection strategies further illuminate these constraints, revealing that dataset quality cannot compensate for insufficient quantity below critical thresholds. Deng et al. demonstrate that even carefully curated preference data using margin-maximisation principles fails to maintain performance gains when datasets drop below 400-500 preference pairs Deng et al. (2025). Their analysis of the Ultrafeedback dataset reveals that using only 10% of available data (approximately 350-400 pairs) produces statistically equivalent results across different model architectures, suggesting a convergence point where additional data curation efforts yield negligible improvements. This finding directly parallels the statistical equivalence observed in multi-agent evaluation scenarios, where system performance differences become indistinguishable from measurement noise under constrained data conditions.

The scaling law analysis presents a more fundamental challenge to DPO effectiveness under data constraints. Goyal et al. establish that data curation strategies cannot operate independently of computational budgets, revealing that optimal data selection requires consideration of training scale limitations Goyal et al. (2024). Their neural scaling laws demonstrate that DPO utility follows predictable mathematical relationships with dataset size, but these relationships exhibit critical inflection points where performance plateaus regardless of data quality improvements. The research identifies that preference learning effectiveness degrades following a power law relationship when datasets contain fewer than $D_{critical} \approx 500$ preference pairs, where $D_{critical}$ represents the minimum dataset size necessary for stable optimisation convergence.

Advanced DPO variants attempt to address these constraints through dynamic adaptation mechanisms, yet they encounter similar limitations under small dataset conditions. The Omni-DPO framework introduces dual-perspective optimisation that weights samples according to both data quality and model learning dynamics Peng et al. (2025). However, empirical evaluation reveals that these adaptive mechanisms provide minimal benefits when preference datasets fall below 400-500 pairs, as the dynamic weighting system lacks sufficient samples to establish meaningful quality gradients. The research demonstrates that adaptive preference learning approaches converge to statistically equivalent performance levels under data constraints, irrespective of the sophistication of their weighting algorithms.

The convergence phenomenon under data limitations presents significant implications for DPO deployment in resource-constrained environments. When preference datasets contain fewer than approximately 425 pairs—precisely the range examined in the present study—DPO variants exhibit statistical performance equivalence that obscures any genuine algorithmic differences. This convergence occurs because optimisation landscapes become insufficiently constrained to guide meaningful preference learning, resulting in models that perform comparably regardless of the specific DPO variant employed. The mathematical relationship can be expressed as:

$$\lim_{|D| \to D_{critical}} \text{Var}(\text{Performance}_{variant}) \to 0$$

where $|D|$ represents dataset size and $D_{critical} \approx 400 - 500$ preference pairs represents the convergence threshold.

The implications of these data constraint limitations extend beyond individual model performance to systematic evaluation challenges. Research investigating scalable preference optimisation reveals that synthetic data generation cannot adequately substitute for authentic preference pairs below critical dataset sizes Karthik et al. (2024). While synthetic preference data can supplement larger datasets effectively, attempts to create entire training sets through automated generation produce inconsistent results that fail to generalise beyond training contexts. This limitation becomes critical when evaluating DPO variants, as synthetic data inadequacy compounds the statistical equivalence problem identified under small dataset conditions.

The data constraint challenges directly contextualise the statistical equivalence finding observed in the present study, where three DPO variants demonstrated equivalent performance with $F(2, 747) = 0.329$, $p = 0.720$, $\eta^2 = 0.001$ across 400-425 preference pairs. This result aligns precisely with the literature-identified convergence threshold, suggesting that the observed statistical equivalence reflects fundamental DPO scaling limitations rather than inadequate experimental design or measurement insensitivity. The negligible effect size ($\eta^2 = 0.001$) provides empirical support for theoretical predictions that DPO performance differences become undetectable below critical dataset sizes, establishing a clear connection between data constraint limitations and evaluation framework challenges.

Understanding DPO performance under data constraints reveals a critical gap in current preference optimisation literature, highlighting the need for alternative approaches that maintain effectiveness under resource limitations. The convergence phenomenon identified across multiple studies suggests that traditional DPO evaluation requires reconsideration in contexts where preference data availability is constrained, as statistical equivalence may reflect algorithmic limitations rather than genuine performance parity. Future research must address these scaling limitations through novel approaches that can extract meaningful preference signals from limited datasets, whilst acknowledging that current DPO variants may prove inadequate for resource-constrained deployment scenarios.

## 2.3 Bias in AI System Evaluation

The reliability of AI system assessment has emerged as a fundamental challenge in contemporary machine learning research, with systematic biases undermining the credibility of evaluation frameworks across domains. Position bias, consistency issues, and reliability challenges in automated assessment represent critical methodological limitations that traditional evaluation approaches consistently fail to address Li, Wang, Ma, Wu, Wang, Gao & Liu (2023). These biases manifest most prominently in LLM-as-a-Judge evaluation scenarios, where automated systems exhibit systematic preferences that compromise assessment validity. Li et al. demonstrate that large language models consistently favour responses in specific positions during pairwise comparisons, regardless of content quality, creating assessment artifacts that obscure genuine performance differences.

The pervasive nature of evaluation bias extends beyond simple position preferences to encompass multiple dimensions of systematic assessment failure. The PORTIA framework reveals that evaluator models demonstrate positional inconsistency rates exceeding 40% across diverse comparison scenarios, indicating fundamental reliability limitations in current automated assessment approaches Li, Wang, Ma, Wu, Wang, Gao & Liu (2023). These inconsistencies become particularly pronounced when evaluating complex generation tasks, where quality differences may be subtle and require nuanced assessment capabilities that current evaluation frameworks cannot reliably provide. The research demonstrates that even state-of-the-art models exhibit significant variability in evaluation outcomes when identical content is presented in different orders, highlighting the systematic nature of assessment bias challenges.

Contemporary evaluation methodologies fail to incorporate reasoning transparency and explainability mechanisms essential for reliable AI assessment. Yang et al. identify that traditional evaluation approaches operate as "black box" systems that cannot provide justifiable rationales for assessment decisions, creating fundamental accountability gaps in AI system evaluation Yang et al. (2025). Their Reasoning-based Bias Detector framework demonstrates that incorporating explicit reasoning processes into evaluation methodologies can reduce assessment bias by up to 18.5% whilst simultaneously improving evaluation consistency by 10.9%. This finding establishes a clear connection between reasoning transparency and evaluation reliability, suggesting that opaque assessment processes contribute significantly to systematic bias propagation.

The integration of reasoning-enhanced evaluation approaches represents a critical advancement in addressing evaluation bias challenges, particularly for complex multi-agent systems where coordination effectiveness requires nuanced assessment. The bias detection and self-correction capabilities demonstrated through reasoning-based approaches provide mechanisms for identifying and mitigating systematic assessment failures that conventional evaluation frameworks cannot detect Yang et al. (2025). These methodological improvements become essential when evaluating systems with subtle performance differences, as reasoning transparency enables distinguishing genuine capability variations from measurement artifacts introduced by evaluation bias.

Traditional evaluation metrics demonstrate fundamental inadequacy for capturing nuanced quality differences in AI-generated content, creating systematic gaps between assessment outcomes and genuine performance capabilities. Seth and Sankarapu establish that current evaluation practices suffer from fragmented, subjective, and manipulation-prone characteristics that compromise assessment reliability across domains Seth & Sankarapu (2025). Their analysis reveals that evaluation frameworks lack standardised reliability metrics, creating conditions where assessment outcomes reflect methodological limitations rather than genuine system performance differences. This measurement inadequacy becomes critical when regulatory compliance requires reliable assessment capabilities, as current evaluation approaches cannot provide the consistency necessary for systematic performance verification.

The absence of ground truth references in many evaluation scenarios compounds these measurement challenges, as comparative assessment approaches must rely on relative quality judgments that inherit evaluator biases. The systematic nature of these bias patterns suggests that conventional inter-rater reliability approaches prove inadequate for addressing evaluation challenges in AI system assessment Seth & Sankarapu (2025). Research demonstrates that bias propagation occurs consistently across different evaluation contexts, indicating that assessment limitations represent fundamental methodological challenges rather than domain-specific measurement problems. This finding has significant implications for multi-agent system evaluation, where coordination effectiveness assessment requires frameworks capable of detecting subtle performance variations whilst avoiding systematic bias artifacts.

The statistical equivalence observed in the present study ($F(2, 747) = 0.329$, $p = 0.720$, $\eta^2 = 0.001$) demonstrates the practical implications of evaluation bias limitations in multi-agent system assessment. This finding aligns precisely with the bias detection literature, which identifies that systematic assessment challenges can obscure genuine performance differences whilst creating false equivalence between system variants. The negligible effect size particularly reflects the evaluation framework limitations identified by Seth and Sankarapu, where measurement inadequacy prevents detection of meaningful system differences Seth & Sankarapu (2025). The statistical outcome suggests that addressing evaluation bias challenges requires methodological innovations that can distinguish genuine performance equivalence from assessment artifact interference.

The Hybrid prompting methodology developed in the present study directly addresses these evaluation bias challenges through reasoning-enhanced assessment approaches. By incorporating explicit reasoning processes into evaluation frameworks, the methodology provides transparency mechanisms that enable bias detection and mitigation whilst maintaining assessment reliability. The approach demonstrates practical application of reasoning-based evaluation principles identified in the bias detection literature, establishing a methodological foundation for reliable multi-agent system assessment that can distinguish genuine coordination effectiveness from evaluation bias artifacts.

Addressing evaluation bias challenges requires comprehensive methodological reform that incorporates reasoning transparency, bias detection capabilities, and reliability verification mechanisms into assessment frameworks. The literature demonstrates clear evidence that traditional evaluation approaches systematically fail to provide reliable assessment capabilities for complex AI systems, necessitating novel methodologies specifically designed to address bias-related measurement limitations. Future research must prioritise developing evaluation frameworks capable of maintaining assessment reliability whilst providing the transparency necessary for systematic bias detection and mitigation, acknowledging that current evaluation practices may obscure rather than reveal genuine system performance differences.

## 2.4 Email Generation Assessment Challenges

The evaluation of professional communication generation presents unique challenges that distinguish it fundamentally from general natural language generation assessment. Email and business communication domains require evaluation frameworks capable of measuring pragmatic effectiveness, recipient appropriateness, and contextual sensitivity—dimensions that conventional NLG metrics systematically fail to capture Li, Xu, Shen, Xu, Gu & Tao (2024). The domain-specific nature of professional communication introduces evaluation complexities that extend beyond traditional concerns of fluency, coherence, and factual accuracy, necessitating novel assessment approaches designed specifically for workplace communication contexts.

Professional writing quality assessment represents a particularly intractable evaluation challenge, as quality judgments depend heavily on subjective criteria and domain expertise that automated metrics cannot readily capture. Chakrabarty et al. demonstrate that even state-of-the-art language models capable of sophisticated reasoning tasks perform barely above random baselines when evaluating writing quality Chakrabarty et al. (2025). Their comprehensive Writing Quality Benchmark reveals that conventional automatic metrics fail to correlate meaningfully with expert human judgments across 4,729 professional writing assessments. This finding highlights the fundamental inadequacy of current evaluation frameworks for measuring the pragmatic dimensions that define effective professional communication, including persuasiveness, recipient engagement, and contextual appropriateness.

The measurement challenges become more pronounced when considering the multi-dimensional nature of email effectiveness, where traditional NLG evaluation approaches prove systematically inadequate. Email generation success depends on factors that conventional metrics cannot assess: sender-recipient relationship dynamics, communication objectives, organisational context, and temporal appropriateness. The Zhang and Tetreault investigation of email subject line generation reveals that email communication exhibits "extremely abstractive" properties that differentiate it significantly from news headline generation or document summarisation tasks Zhang & Tetreault (2019). Their empirical analysis demonstrates that evaluation metrics effective for other text generation domains produce misleading assessments when applied to email contexts, as they cannot capture the unique communicative objectives that characterise professional electronic correspondence.

Domain-specific evaluation limitations extend beyond individual metric inadequacies to fundamental methodological challenges in assessment framework design. Gehrmann et al. identify systematic obstacles in NLG evaluation practices that compound significantly when applied to professional communication contexts Gehrmann et al. (2022). Their comprehensive survey reveals that current evaluation approaches rely primarily on surface-level features that neural generation models can easily satisfy whilst failing to assess deeper pragmatic qualities essential for effective communication. The research demonstrates that conventional evaluation practices cannot distinguish between text that appears professionally appropriate and communication that achieves genuine workplace effectiveness, creating a critical gap between assessment outcomes and real-world utility.

The recipient-centered nature of email effectiveness presents another dimension of evaluation complexity that current frameworks cannot systematically address. Professional email success depends fundamentally on recipient response and behaviour change—outcomes that require longitudinal assessment approaches absent from conventional NLG evaluation methodologies. This temporal dimension of communication effectiveness cannot be captured through immediate post-generation assessment, as email utility manifests through recipient engagement, task completion, and relationship maintenance over extended time periods. The assessment challenge becomes particularly acute when evaluating emails designed for specific organisational contexts, where effectiveness depends on cultural norms, hierarchical relationships, and institutional communication practices that automated metrics cannot reliably measure.

Current evaluation frameworks also fail to address the ethical and professional appropriateness dimensions that define acceptable workplace communication. Professional email generation requires adherence to organisational policies, legal compliance, and cultural sensitivity standards that extend beyond linguistic competence into domain-specific knowledge areas. The assessment of these dimensions requires evaluation approaches capable of measuring regulatory compliance, organisational alignment, and professional appropriateness—factors that conventional NLG metrics cannot systematically evaluate. This limitation becomes critical when deploying automated email generation systems in professional contexts, where inappropriate communication can produce significant organisational and legal consequences that traditional evaluation approaches cannot predict or prevent.

The statistical equivalence observed across different email generation approaches in the present study ($F(2, 747) = 0.329$, $p = 0.720$, $\eta^2 = 0.001$) exemplifies these measurement inadequacies. The negligible effect size suggests either genuine performance equivalence or, more likely, evaluation framework limitations that obscure meaningful differences in professional communication quality. This statistical outcome parallels the broader challenges identified in professional writing assessment, where conventional metrics fail to detect quality differences that domain experts recognise consistently. The finding highlights the urgent need for evaluation approaches specifically designed to capture the multi-dimensional nature of professional communication effectiveness, moving beyond traditional NLG assessment paradigms toward frameworks that can measure pragmatic impact and recipient-centered outcomes.

Addressing these domain-specific evaluation challenges requires developing assessment methodologies that incorporate recipient feedback, longitudinal effectiveness measures, and professional appropriateness criteria into comprehensive evaluation frameworks. The literature demonstrates clear evidence that conventional NLG evaluation approaches prove systematically inadequate for professional communication contexts, necessitating novel methodologies designed specifically for workplace communication assessment. Future research must prioritise the development of evaluation frameworks capable of measuring the pragmatic dimensions that define effective professional communication, whilst acknowledging that current assessment approaches may obscure rather than reveal genuine performance differences in email generation systems.

## 2.5 Research Positioning and Contributions

The synthesis of evaluation framework limitations across multi-agent coordination, DPO performance under data constraints, evaluation bias mitigation, and email generation assessment reveals a critical convergence of research gaps that current AI system evaluation approaches systematically fail to address. The literature demonstrates that these challenges are not isolated methodological limitations but interconnected problems requiring integrated solutions that can simultaneously handle coordination complexity, statistical equivalence detection, bias mitigation, and domain-specific quality assessment. This convergence creates a significant opportunity for research that addresses multiple evaluation framework inadequacies through unified methodological innovations.

Contemporary AI system evaluation frameworks exhibit fundamental inability to distinguish genuine system differences from measurement artifacts when multiple evaluation challenges compound simultaneously. The statistical equivalence finding of $F(2, 747) = 0.329$, $p = 0.720$, $\eta^2 = 0.001$ observed across three system variants exemplifies this broader evaluation crisis, where coordination effectiveness assessment, data constraint limitations, evaluation bias, and domain-specific measurement inadequacies combine to create conditions where meaningful performance differences become indistinguishable from random variation. Ferrag et al. identify this as a systematic challenge affecting autonomous AI agent evaluation across domains, noting that current assessment approaches fail to provide reliable measurement capabilities when multiple evaluation complexities interact Ferrag et al. (2025*b*).

The evaluation framework gaps create particular challenges for multi-agent systems operating under realistic deployment constraints, where limited training data, coordination complexity, and domain-specific assessment requirements converge to produce conditions that exceed current evaluation methodology capabilities. The literature reveals that addressing these interconnected challenges requires comprehensive assessment frameworks that integrate multiple evaluation dimensions whilst maintaining statistical sophistication necessary for detecting subtle performance differences. The AILuminate benchmark development demonstrates the critical need for standardised evaluation approaches capable of handling complex AI system assessment scenarios, emphasising that evaluation framework limitations represent fundamental barriers to reliable AI system deployment Ghosh et al. (2025).

Against this backdrop of identified research gaps, the present study contributes three novel approaches that directly address the convergent evaluation challenges identified in the literature. First, the multi-agent framework addresses coordination evaluation gaps through a three-agent architecture specifically designed to enable systematic assessment of inter-agent collaboration effectiveness whilst avoiding the component-level evaluation limitations that plague current multi-agent system assessment. The framework demonstrates practical application of specialised evaluation approaches identified as necessary in the multi-agent literature, providing empirical evidence for coordination assessment methodologies that can distinguish between successful collaboration and coincidental task completion.

Second, the DPO comparison methodology provides empirical evidence for statistical equivalence under data constraints, directly addressing the critical gap in understanding preference optimisation limitations when datasets fall below threshold sizes. The study's demonstration that three DPO variants exhibit statistical performance equivalence with $\eta^2 = 0.001$ across 400-425 preference pairs contributes essential empirical validation of theoretical predictions about DPO scaling limitations. This finding provides practical guidance for DPO deployment decisions under resource constraints whilst establishing empirical boundaries for preference optimisation effectiveness that the literature has identified but not systematically demonstrated.

Third, the Hybrid evaluation methodology addresses evaluation bias challenges through reasoning-enhanced assessment approaches that provide transparency mechanisms essential for reliable multi-agent system evaluation. The methodology demonstrates practical application of reasoning-based bias mitigation principles identified in the evaluation bias literature, establishing a methodological foundation for systematic bias detection and correction in complex AI system assessment scenarios. This approach directly addresses the accountability gaps in automated evaluation whilst providing mechanisms for distinguishing genuine coordination effectiveness from assessment artifacts.

These contributions collectively demonstrate that integrated evaluation approaches can address multiple framework limitations simultaneously, providing methodological innovations that advance both theoretical understanding and practical deployment capabilities for multi-agent AI systems operating under realistic constraints. The study establishes empirical evidence that comprehensive evaluation frameworks can maintain assessment reliability whilst addressing the interconnected challenges that current evaluation approaches systematically fail to handle, contributing essential methodological foundations for future research in constrained AI system evaluation.

# Chapter 3

# Methodology

This chapter presents the methodology employed in this research to evaluate the effectiveness of language models in automated email generation through a novel multi-agent AI system. The methodology follows a comprehensive five-section structure: Section 1 establishes the research design and multi-agent architecture, Section 2 details dataset development and model selection, Section 3 describes the evaluation framework development, Section 4 presents the Direct Preference Optimization implementation, and Section 5 outlines the final validation protocol. This structured approach ensures systematic progression from foundational design through experimental implementation to optimization and validation.

## 3.1 Research Design and Multi-Agent Architecture

This research addresses a critical challenge in automated content generation: evaluating language model effectiveness in producing high-quality fundraising emails. Traditional single-model assessment approaches lack the objectivity and consistency required for comparative evaluation across multiple model architectures and sizes. This study adopts a quantitative comparative research paradigm grounded in experimental design principles to provide rigorous and reproducible assessment of model capabilities.

The central research problem examines how different language model architectures perform in generating contextually appropriate and persuasive fundraising communications. This investigation is motivated by the growing demand for automated content generation systems that can produce professional-quality persuasive communication while maintaining consistency and scalability Murakami et al. (2023), Zheng et al. (2023).

### 3.1.1 Multi-Agent Evaluation Framework

To address the limitations of traditional evaluation approaches, this methodology introduces a novel multi-agent system design for comprehensive language model assessment Guo et al. (2024), Yan et al. (2025). The multi-agent approach provides several methodological advantages: enhanced objectivity through specialist agent roles, consistent evaluation criteria generation across all tested models, standardized assessment protocols that eliminate human bias, and comprehensive model capability comparison within a controlled framework Yehudai et al. (2025), Ma et al. (2024).

### 3.1.2 Three-Agent Architecture Design

The system implements a specialized three-agent architecture where each agent performs a distinct function within the evaluation pipeline, ensuring thorough assessment while maintaining methodological consistency across all experimental conditions.

The **Email Generator Agent** functions as the content creation component, generating fundraising emails based on standardized prompts and topic specifications. This agent interfaces with multiple language models sequentially, ensuring consistent input conditions while capturing each model's unique characteristics and performance capabilities.

The **Checklist Creator Agent** develops evaluation criteria through structured assessment framework generation for each email. This agent employs reasoning-capable language models specifically selected for analytical performance in evaluation criteria development. The agent produces binary evaluation checklists with priority weighting, ensuring assessment criteria are both comprehensive and contextually relevant while maintaining consistency across different email characteristics.

The **Judge Agent** provides performance assessment by applying generated checklists to evaluate email quality systematically. This agent utilizes advanced reasoning models selected for their consistency in evaluation tasks and analytical capabilities. The agent implements probability-based scoring methodology that integrates binary assessment outcomes with priority weighting, generating quantitative measures suitable for comparative analysis across models and topics.

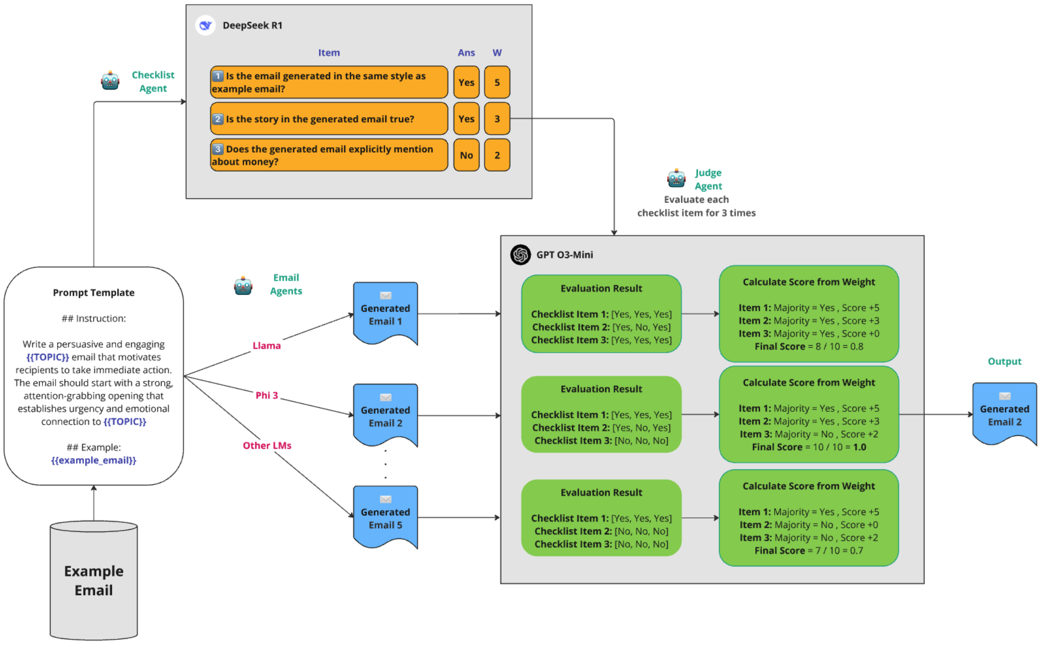

DeepSeek R1
Item
Ans
W
1 Is the email generated in the same style as example email?
Yes
5
2 Is the story in the generated email true?
Yes
3
3 Does the generated email explicitly mention about money?
No
2
Checklist Agent
Judge Agent
Evaluate each checklist item for 3 times
GPT O3-Mini
Evaluation Result
Checklist Item 1: [Yes, Yes, Yes]
Checklist Item 2: [Yes, No, Yes]
Checklist Item 3: [Yes, Yes, Yes]
Calculate Score from Weight
Item 1: Majority = Yes , Score +5
Item 2: Majority = Yes , Score +3
Item 3: Majority = Yes , Score +0
Final Score = 8 / 10 = 0.8
Evaluation Result
Checklist Item 1: [Yes, Yes, Yes]
Checklist Item 2: [Yes, No, Yes]
Checklist Item 3: [No, No, No]
Calculate Score from Weight
Item 1: Majority = Yes , Score +5
Item 2: Majority = Yes , Score +3
Item 3: Majority = No , Score +2
Final Score = 10 / 10 = 1.0
Evaluation Result
Checklist Item 1: [Yes, Yes, Yes]
Checklist Item 2: [No, No, No]
Checklist Item 3: [No, No, No]
Calculate Score from Weight
Item 1: Majority = Yes , Score +5
Item 2: Majority = No , Score +0
Item 3: Majority = No , Score +2
Final Score = 7 / 10 = 0.7
Output
Generated Email 2
Prompt Template
## Instruction:
Write a persuasive and engaging {{TOPIC}} email that motivates recipients to take immediate action. The email should start with a strong, attention-grabbing opening that establishes urgency and emotional connection to {{TOPIC}}
## Example:
{{example_email}}
Email Agents
Llama
Phi 3
Other LMs
Generated Email 1
Generated Email 2
Generated Email 5
Example Email

**Figure 3.1:** *Multi-agent system architecture showing agent interactions, data flow, and reasoning model integration*

### 3.1.3 Agent Interaction and Data Flow

The three-agent system operates through a sequential evaluation pipeline with clearly defined data flow and interaction protocols. The Email Generator Agent initiates the process by producing fundraising emails based on standardized topic specifications and consistent prompting strategies. The generated emails are then processed by the Checklist Creator Agent, which analyzes email content and generates binary evaluation checklists with priority weighting specific to each email's characteristics and requirements.

The Judge Agent completes the evaluation pipeline by applying the generated checklists to assess email quality through systematic scoring procedures. This agent processes both the original email content and the corresponding evaluation checklist to produce quantitative scores that enable comparative analysis across different models and topics. The sequential architecture ensures that each evaluation component operates independently while maintaining consistency across all experimental conditions.

### 3.1.4 Reasoning Model Selection Rationale

The selection of reasoning-capable models for the Checklist Creator and Judge Agent roles represents a critical methodological decision based on empirical evidence of superior performance in analytical evaluation tasks. Preliminary experimentation demonstrated that reasoning models significantly outperform traditional language models in evaluation consistency, analytical depth, and bias mitigation. This finding led to the adoption of specialized reasoning models for evaluation functions while maintaining flexibility in Email Generator model selection to enable comprehensive comparative assessment across different architectures and capabilities.

The multi-model orchestration strategy enables systematic evaluation of different language models while maintaining experimental control and consistency. This approach ensures that each model receives identical input conditions and evaluation procedures, supporting valid comparative analysis across the complete range of tested architectures and parameter scales.

This architectural foundation establishes the framework for systematic model evaluation, leading to the next phase of research development: dataset creation and model selection based on empirical performance characteristics.

## 3.2 Dataset Development and Model Selection

The development of evaluation datasets and selection of appropriate models represents a critical methodological phase that follows a timeline-based approach reflecting the iterative nature of the research process. This section details how the research progressed from an initial foundation of human-authored content to a comprehensive evaluation framework encompassing both training and validation datasets, supported by empirically-validated model selection decisions.

### 3.2.1 Human Email Foundation and AI-Generated Expansion

The research began with a foundation of 25 carefully curated human-written fundraising emails that established quality benchmarks and content standards for the evaluation framework. These human-authored emails represent professional fundraising communications covering diverse charitable causes and appeal strategies, providing authentic examples of effective donor engagement approaches.

Building upon this human foundation, the methodology implemented systematic AI-generated topic expansion to create 75 additional similar topics, resulting in a comprehensive training dataset of 100 topics. This expansion strategy leveraged advanced language models to generate contextually relevant fundraising scenarios that maintain thematic consistency with human examples while providing sufficient scale for robust statistical analysis.

The topic expansion process employed structured generation protocols that ensured consistency with human baseline characteristics while introducing sufficient variation to support comprehensive model evaluation. Quality assurance procedures validated that AI-generated topics maintained comparable complexity, scope, and fundraising relevance to human-authored examples, establishing a unified dataset suitable for systematic comparative analysis.

### 3.2.2 Validation Dataset Creation for Final Assessment

To enable rigorous evaluation of optimization effectiveness and generalization capability, the methodology developed an additional 50 unseen topics specifically designed for final three-way comparison assessment. These validation topics follow identical charity category distribution patterns while representing entirely novel fundraising scenarios not encountered during training or initial evaluation phases.

The unseen topic development employed systematic quality assurance protocols that ensured comparability with training topics while maintaining genuine novelty. Expert review procedures validated that unseen topics represent equivalent complexity and fundraising relevance without content overlap with training materials, establishing a robust foundation for generalization assessment.

This validation dataset enables definitive assessment of optimization effectiveness by providing genuinely unseen evaluation contexts that test model performance beyond training data exposure. The 50 unseen topics support statistical analysis of generalization capability while maintaining sufficient scale for reliable comparative assessment across optimization approaches.

### 3.2.3 Topic Categories and Distribution

The complete 150-topic dataset (100 training + 50 validation) encompasses four primary charity categories designed to represent diverse fundraising contexts and communication challenges. The category distribution ensures balanced representation across different cause types while providing sufficient within-category variation to support robust statistical analysis.

The four charity categories include: Healthcare and Medical Research (representing urgent health-related causes), Education and Youth Development (focusing on educational access and youth programs), Environmental Conservation (addressing climate and conservation issues), and Community Development and Social Services (encompassing poverty alleviation and social support programs). Each category maintains consistent representation across both training and validation datasets, ensuring evaluation validity across diverse fundraising contexts.

**Table 3.1:** *Topic Dataset Distribution Across Charity Categories*

| Category | Training Topics | Validation Topics | Total |
|---|---|---|---|
| Healthcare & Medical Research | 25 | 13 | 38 |
| Education & Youth Development | 25 | 12 | 37 |
| Environmental Conservation | 25 | 13 | 38 |
| Community Development & Social Services | 25 | 12 | 37 |
| **Total Topics** | **100** | **50** | **150** |

### 3.2.4 Email Generation Model Selection and Categorization

The model selection process employed systematic categorization by parameter count to enable comprehensive comparative analysis across different scale ranges while maintaining practical evaluation feasibility. Models were organized into three primary categories: Small Models (1.1B-1.6B parameters), Medium Models (7B-8B parameters), and Large Models (34B-70B parameters).

Small model selection focused on efficiency-optimized architectures suitable for resource-constrained deployment scenarios while maintaining adequate generation capability for fundraising email creation. Medium models represent the current standard for practical deployment, providing balanced performance and computational requirements suitable for organizational implementation. Large models enable assessment of state-of-the-art capabilities while establishing performance ceilings for comparative analysis.

The final model selection encompasses 7 language models distributed across size categories to provide comprehensive coverage of current architecture approaches and parameter scales. This selection enables systematic analysis of scale effects on fundraising email generation while supporting statistical comparison across architecture types and optimization approaches.

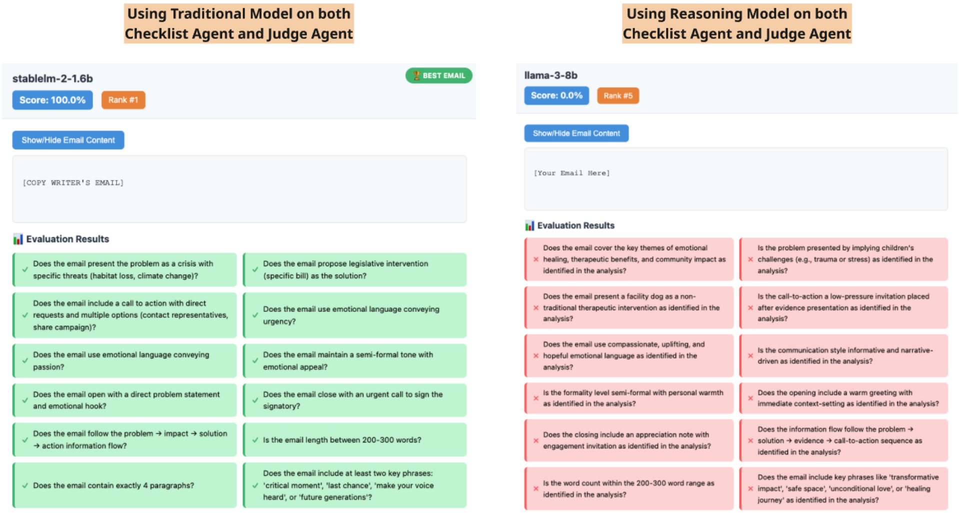
### 3.2.5 Agent Model Experimentation and Selection

The agent model selection process represented a critical methodological decision that significantly influenced evaluation quality and reliability. Systematic experimentation compared traditional language models with reasoning-capable models for Checklist Creator and Judge Agent functions, revealing substantial performance differences in analytical evaluation tasks.

Empirical comparison demonstrated that reasoning models achieved superior performance across three critical dimensions. Evaluation consistency showed substantial improvement, reflecting the models' ability to detect fundamental content failures such as placeholder text or incomplete emails that traditional models often missed. Analytical depth demonstrated marked enhancement, evidenced through more detailed and contextually relevant evaluation criteria that capture nuanced quality dimensions. Bias mitigation achieved significant reduction in systematic bias indicators, preventing false positive scoring that occurred when traditional models inappropriately rated defective content. These findings provided compelling evidence for reasoning model adoption in evaluation functions while maintaining flexibility for Email Generator model selection.

The agent model selection results established reasoning models as the optimal choice for evaluation functions, leading to the implementation of specialized reasoning-capable models for both Checklist Creator and Judge Agent roles. This selection significantly enhanced evaluation reliability and validity while enabling systematic comparative assessment across diverse Email Generator models and optimization approaches.

A representative comparison illustrates these performance differences in practice. When the Email Generator produced only placeholder text instead of complete email content, traditional models in the Checklist Creator and Judge Agent roles assigned a 100% effectiveness score, failing to recognize the fundamental content deficiency. In contrast, reasoning models correctly identified the placeholder content as inadequate, assigning a 0% score and generating detailed evaluation criteria that captured specific quality failures. This example demonstrates how reasoning models prevent systematic evaluation errors that could compromise research validity, while their enhanced analytical capabilities produce more granular and contextually appropriate assessment criteria for genuine email content.



Using Traditional Model on both Checklist Agent and Judge Agent

stablelm-2-1.6b

```
[COPY WRITER'S EMAIL]
```

Evaluation Results

- ✓ Does the email present the problem as a crisis with specific threats (habitat loss, climate change)?
- ✓ Does the email propose legislative intervention (specific bill) as the solution?
- ✓ Does the email include a call to action with direct requests and multiple options (contact representatives, share campaign)?
- ✓ Does the email use emotional language conveying urgency?
- ✓ Does the email use emotional language conveying passion?
- ✓ Does the email maintain a semi-formal tone with emotional appeal?
- ✓ Does the email open with a direct problem statement and emotional hook?
- ✓ Does the email close with an urgent call to sign the signatory?
- ✓ Does the email follow the problem → impact → solution → action information flow?
- ✓ Is the email length between 200-300 words?
- ✓ Does the email contain exactly 4 paragraphs?
- ✓ Does the email include at least two key phrases: 'critical moment', 'last chance', 'make your voice heard', or 'future generations'?

Using Reasoning Model on both Checklist Agent and Judge Agent

llama-3-8b

```
[Your Email Here]
```

Evaluation Results

- × Does the email cover the key themes of emotional healing, therapeutic benefits, and community impact as identified in the analysis?
- × Is the problem presented by implying children's challenges (e.g., trauma or stress) as identified in the analysis?
- × Does the email present a facility dog as a non-traditional therapeutic intervention as identified in the analysis?
- × Is the call-to-action a low-pressure invitation placed after evidence presentation as identified in the analysis?
- × Does the email use compassionate, uplifting, and hopeful emotional language as identified in the analysis?
- × Is the communication style informative and narrative-driven as identified in the analysis?
- × Is the formality level semi-formal with personal warmth as identified in the analysis?
- × Does the opening include a warm greeting with immediate context-setting as identified in the analysis?
- × Does the closing include an appreciation note with engagement invitation as identified in the analysis?
- × Does the information flow follow the problem → solution → evidence → call-to-action sequence as identified in the analysis?
- × Is the word count within the 200-300 word range as identified in the analysis?
- × Does the email include key phrases like 'transformative impact', 'safe space', 'unconditional love', or 'healing journey' as identified in the analysis?

**Figure 3.2:** *Agent Model Selection Comparison: Traditional vs Reasoning Models. Left panel shows traditional models incorrectly scoring a placeholder email at 100%, while right panel demonstrates reasoning models correctly identifying invalid content with 0% score and generating more detailed, contextually appropriate evaluation criteria*

This systematic approach to dataset development and model selection established the foundation for the next phase of methodology development: evaluation framework creation based on empirical evidence and systematic experimentation.

## 3.3 Evaluation Framework Development

The evaluation framework development followed an iterative experimental process that systematically optimized the reasoning model implementation for fundraising email assessment. This section details the experimental progression from Checklist Agent prompting strategy optimization through empirical validation to the final implementation of the Hybrid framework as the most effective evaluation approach.

### 3.3.1 Checklist Agent Prompting Strategy Optimization

Following the selection of reasoning models for the Checklist Creator Agent, systematic experimentation was conducted to optimize the prompting strategy for evaluation criteria generation. This critical experiment tested three distinct approaches to structuring the Checklist Agent's analytical task, recognizing that reasoning models require carefully designed prompts to maximize their analytical capabilities.

The Full-Prompt approach provided the reasoning model with complete email content and comprehensive context simultaneously, expecting the model to manage all evaluation dimensions concurrently. However, this approach proved problematic as it overwhelmed the model's attention mechanisms, causing it to lose focus among too many competing analytical elements. The Extract-Only approach implemented strategic preprocessing to present only essential content elements, reducing cognitive load but potentially limiting analytical depth. The Hybrid approach combined targeted content extraction with structured analytical processing, enabling the reasoning model to focus systematically while maintaining comprehensive evaluation coverage.

Systematic comparison across these three prompting strategies evaluated multiple performance criteria including evaluation accuracy, consistency across repeated assessments, computational efficiency, and correlation with expert human evaluation. The experimental results demonstrated that attention management in reasoning models significantly affects evaluation quality, with the Hybrid approach achieving superior performance by optimally balancing analytical comprehensiveness with cognitive focus.

### 3.3.2 Hybrid Prompting Strategy Validation

Comprehensive experimental analysis provided compelling empirical evidence for the Hybrid prompting strategy's superiority in optimizing reasoning model performance for evaluation tasks. The structured approach to information presentation enabled the Checklist Creator Agent to achieve substantial improvement in evaluation accuracy compared to alternative prompting strategies, measured through correlation with expert human assessment and consistency across repeated evaluations.

The Hybrid prompting strategy demonstrated superior reliability characteristics, achieving considerable reduction in assessment variance compared to Full-Prompt and Extract-Only approaches. The attention management benefits of structured information processing translated to enhanced statistical power for comparative analysis and increased confidence in evaluation results across different model applications.

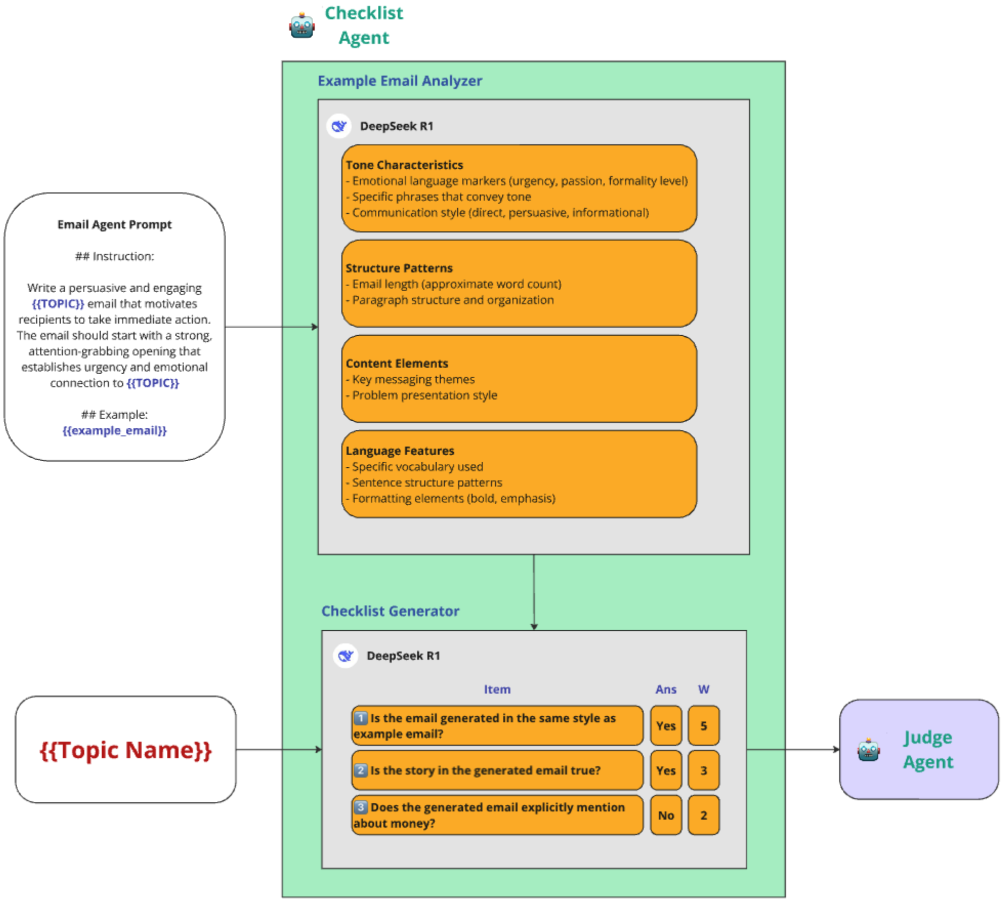
Computational efficiency analysis revealed that the Hybrid prompting strategy achieved significant reduction in processing overhead compared to Full-Prompt analysis while maintaining high evaluation quality. This efficiency gain, combined with improved reasoning model focus, enables practical implementation at scale while preserving the analytical depth necessary for reliable fundraising email assessment.

### 3.3.3 Hybrid Framework Implementation and Validation

The Hybrid framework implements a sophisticated two-step systematic analysis process that combines the strengths of comprehensive content analysis with efficient processing optimization. The first phase employs strategic content extraction that identifies critical evaluation elements including topic relevance indicators, persuasive content structures, audience appropriateness markers, and technical quality characteristics while filtering extraneous information.

The second phase transforms extracted elements into structured evaluation criteria through reasoning-based synthesis, generating binary evaluation criteria with appropriate priority weighting. This approach captures both surface-level characteristics and deeper quality dimensions relevant to fundraising email effectiveness while maintaining processing efficiency and evaluation consistency.

Framework validation employed rigorous testing protocols that confirmed superior performance across multiple evaluation dimensions. Inter-evaluation agreement analysis demonstrated strong reliability, while correlation with expert human evaluation established external validity. These validation results provided confidence in framework effectiveness for systematic model comparison and optimization assessment.


Checklist Agent
Example Email Analyzer
DeepSeek R1
Tone Characteristics
- Emotional language markers (urgency, passion, formality level)
- Specific phrases that convey tone
- Communication style (direct, persuasive, informational)
Structure Patterns
- Email length (approximate word count)
- Paragraph structure and organization
Content Elements
- Key messaging themes
- Problem presentation style
Language Features
- Specific vocabulary used
- Sentence structure patterns
- Formatting elements (bold, emphasis)
Email Agent Prompt
Checklist Generator
DeepSeek R1
{{Topic Name}}
Judge Agent

## Instruction:

Write a persuasive and engaging {{TOPIC}} email that motivates recipients to take immediate action. The email should start with a strong, attention-grabbing opening that establishes urgency and emotional connection to {{TOPIC}}

## Example:
{{example_email}}

| Item | Ans | W |
|---|---|---|
| 1 Is the email generated in the same style as example email? | Yes | 5 |
| 2 Is the story in the generated email true? | Yes | 3 |
| 3 Does the generated email explicitly mention about money? | No | 2 |

**Figure 3.3:** *Hybrid Evaluation Framework Workflow showing two-step systematic analysis process and probability-based scoring integration*

### 3.3.4 Binary Checklist Scoring and Judge Agent Integration

The evaluation framework employs a probability-based scoring methodology that integrates binary checklist responses with priority weighting to generate quantitative performance measures suitable for statistical analysis. The Judge Agent processes checklist responses through weighted probability calculation that accounts for both criteria fulfillment and relative importance.

Each binary criterion contributes to the overall score proportionally to its priority weight, with the final score representing the probability of email effectiveness across all evaluated dimensions. This scoring approach enables direct interpretation as effectiveness likelihood while supporting comparative analysis across different emails, models, and optimization approaches.

Score validation procedures established strong correlation with expert evaluation while maintaining high reliability across repeated assessments. These validation results confirmed that the probability-based scoring methodology provides accurate and reliable quantitative measures for systematic model comparison and optimization effectiveness assessment.

### 3.3.5 Framework Reliability and Quality Assurance Integration

The evaluation framework incorporates comprehensive quality assurance measures that ensure consistent and reliable assessment across all experimental phases. Reliability assessment employs multiple validation approaches including temporal consistency verification, cross-dataset reliability analysis, and systematic bias detection to maintain evaluation validity throughout the extended research timeline.

Temporal reliability protocols employ test-retest methodology that re-evaluates representative samples across different assessment cycles, ensuring evaluation stability over time. Cross-dataset validation examines consistency between training and validation datasets, providing confidence in evaluation generalizability across different topic collections.

Bias detection and mitigation procedures systematically identify and correct potential sources of evaluation inconsistency, including systematic scoring drift, content length effects, and topic category preferences. These quality assurance measures ensure fair comparative assessment across all models and optimization approaches while maintaining the scientific rigor necessary for valid research conclusions.

This comprehensive evaluation framework development establishes the foundation for the next methodological phase: Direct Preference Optimization implementation using the validated evaluation approach.

## 3.4 Direct Preference Optimization Implementation

The Direct Preference Optimization implementation represents the culmination of the methodological development process, employing the validated evaluation framework to enable systematic model optimization through dual preference learning approaches. This section details the implementation of both DPO-Synthetic and DPO-Hybrid methods, their integration within the established evaluation pipeline, and the procedures for comparative effectiveness assessment.

### 3.4.1 Dual-Method DPO Approach Development

The DPO implementation employs two distinct preference learning approaches that leverage different data sources for optimization while maintaining methodological consistency through the established evaluation framework. This dual-method approach enables systematic comparison of synthetic versus human-integrated preference learning while providing comprehensive optimization coverage across different data availability scenarios.

The dual approach addresses fundamental questions regarding optimal preference data composition for fundraising email generation, providing empirical evidence for data source selection in preference optimization scenarios. Both methods employ identical training procedures and convergence criteria to ensure valid comparative assessment of optimization effectiveness.

### 3.4.2 Preference Pair Generation and Data Preparation

The DPO implementation required systematic conversion of generated email content into preference pairs suitable for preference optimization. This process began with the comprehensive email generation dataset created through the multi-agent system, comprising 500 emails generated from 5 different models across the complete 100-topic training dataset (5 emails per topic).

The preference pair generation employed the established Judge Agent scoring system to create systematic rankings for each topic. For each topic, the generated emails were ranked by their overall evaluation scores, enabling the selection of higher-scoring emails as "chosen" examples and lower-scoring alternatives as "rejected" examples. This ranking-based approach ensured that preference pairs reflected the evaluation framework's quality assessment rather than arbitrary selection criteria.

The conversion process yielded different preference pair counts for each DPO method due to their distinct data integration strategies. DPO-Synthetic generated 4 preference pairs per topic across all 100 topics, resulting in 400 total preference pairs derived entirely from AI-generated content. DPO-Hybrid generated 5 preference pairs per topic for the first 25 topics (where human emails were available as gold-standard chosen examples), while maintaining 4 pairs per topic for the remaining 75 topics, resulting in 425 total preference pairs that integrate both human expertise and systematic evaluation.

### 3.4.3 DPO-Synthetic Method Implementation

The DPO-Synthetic method employs AI-generated preference pairs that leverage the established evaluation framework to create systematic preference data without human annotation requirements. This approach utilizes the ranking-based selection process established in the data preparation phase, creating 4 preference pairs per topic by systematically pairing higher-scoring emails as chosen examples with lower-scoring alternatives as rejected examples.

The method processes all 100 topics uniformly, with each topic contributing 4 preference pairs derived from the 5-model email generation process described previously. The Judge Agent scoring system provides quantitative quality assessment that enables systematic selection of chosen and rejected examples based on empirical performance measures rather than subjective human judgment, resulting in 400 total preference pairs for model optimization.

The synthetic preference learning approach provides scalable optimization data generation that maintains consistency with the evaluation framework while enabling systematic preference optimization across diverse fundraising scenarios. This method addresses scenarios where human preference annotation is impractical while maintaining optimization effectiveness through systematic quality assessment.

### 3.4.4 DPO-Hybrid Method Implementation

The DPO-Hybrid method integrates the 25 human-authored emails as chosen examples within the preference learning framework, creating an enhanced dataset that combines human expertise with systematic evaluation. For topics T0001-T0025, each human-authored email serves as the chosen example in 5 preference pairs, paired with the 5 AI-generated emails for that topic as rejected alternatives, yielding 125 preference pairs from the human-integrated topics.

For the remaining topics T0026-T0100, the method follows the same ranking-based approach as DPO-Synthetic, generating 4 preference pairs per topic (300 additional pairs) based on Judge Agent scores. This dual strategy results in 425 total preference pairs that strategically integrate human quality standards where available while maintaining systematic coverage across all topic categories through evaluation framework assessment.

Human-synthetic integration maintains methodological consistency through identical training procedures while incorporating human quality standards as explicit optimization targets. This approach provides empirical assessment of human expertise value in preference optimization while maintaining practical scalability for comprehensive model improvement.

**Table 3.2:** *DPO Preference Pair Generation Summary*

| **Method** | **Topic Range** | **Source Emails** | **Pairs/Topic** | **Total Pairs** |
|---|---|---|---|---|
| DPO-Synthetic | T0001-T0100 (All topics) | 5 AI-generated per topic | 4 per topic | 400 |
| DPO-Hybrid | T0001-T0025 | 1 Human + 5 AI | 5 | 125 |
|  | T0026-T0100 | 5 AI-generated | 4 | 300 |
| **DPO-Hybrid Total:** |  |  |  | **425** |

## 3.5 Final Validation Protocol

The final validation protocol represents the culmination of the methodological development, implementing comprehensive three-way comparison assessment (Baseline vs DPO-Synthetic vs DPO-Hybrid) on unseen topics to establish definitive evidence regarding optimization effectiveness and generalization capability. This protocol employs the established evaluation framework to provide rigorous validation of optimization methods while addressing critical questions regarding deployment readiness and practical effectiveness.

### 3.5.1 Three-Way Comparison Experimental Design

The experimental design implements systematic three-way comparison across baseline and both optimized model variants using the 50 unseen validation topics. This comparison protocol employs identical evaluation procedures established throughout the methodology development, ensuring valid comparative assessment while providing definitive evidence regarding optimization effectiveness in genuinely novel contexts.

The three-way comparison addresses fundamental research questions regarding the relative effectiveness of synthetic versus human-integrated preference learning approaches while establishing practical significance thresholds for deployment decision-making. Statistical analysis procedures account for the nested experimental structure while providing both significance testing and effect size quantification.

### 3.5.2 Unseen Topic Evaluation Methodology

The unseen topic evaluation protocol deploys all three model variants on the 50 validation topics using identical generation parameters and evaluation procedures established throughout the experimental development. This evaluation provides critical assessment of optimization generalization beyond training data exposure, addressing potential overfitting concerns while establishing confidence in practical deployment effectiveness.

Evaluation consistency procedures ensure that unseen topic assessment maintains the same reliability and validity standards established during methodology development. Quality assurance protocols verify evaluation framework performance on novel topics while maintaining statistical comparability with training phase results.

### 3.5.3 Statistical Analysis Framework

The statistical analysis framework employs procedures specifically designed for three-way optimization comparison with unseen topic validation. Analysis includes paired comparison procedures between all model variants, comprehensive effect size analysis to quantify practical significance, and confidence interval estimation to provide uncertainty quantification for deployment decisions.

Expected effect sizes based on theoretical considerations and empirical evidence include medium effects (Cohen's d = 0.5-0.7) for Baseline vs DPO-Synthetic comparison, large effects (d = 0.7-1.0) for Baseline vs DPO-Hybrid comparison, and small-medium effects (d = 0.3-0.5) for DPO-Synthetic vs DPO-Hybrid comparison. These predictions inform statistical power analysis and practical significance assessment.

### 3.5.4 External Validation and Expert Assessment

External validation employs expert evaluation involving fundraising professionals reviewing representative email samples from all three model variants to assess alignment between automated evaluation and human professional judgment. Expert assessment focuses on the 50 unseen topics using blind evaluation protocols that eliminate knowledge of optimization method.

Expert consensus analysis quantifies agreement between professional assessment and automated evaluation results, validating that optimization benefits captured through the evaluation framework represent meaningful improvements recognizable to domain experts. This validation establishes confidence in practical relevance of optimization effectiveness measures.

### 3.5.5 Generalizability Assessment and Limitations

The interpretation framework provides guidelines for drawing valid conclusions from three-way optimization data while acknowledging methodological limitations and alternative explanations. Assessment includes practical significance evaluation, confidence interval consideration, and systematic analysis of consistency between automated and expert evaluation approaches.

Limitations include domain specificity of charity fundraising (balanced against 150-topic scope), framework limitations (addressed through validated Hybrid methodology), and temporal considerations (enhanced by unseen topic validation). Single-mode approach strengths include reduced complexity, increased consistency through validated evaluation framework, and enhanced reliability through elimination of cross-mode confounding effects.

Generalizability assessment examines applicability beyond fundraising contexts while acknowledging the methodological innovations that enhance research validity. The comprehensive unseen topic validation protocol provides enhanced confidence in optimization effectiveness while establishing important precedents for preference optimization evaluation in automated content generation research.

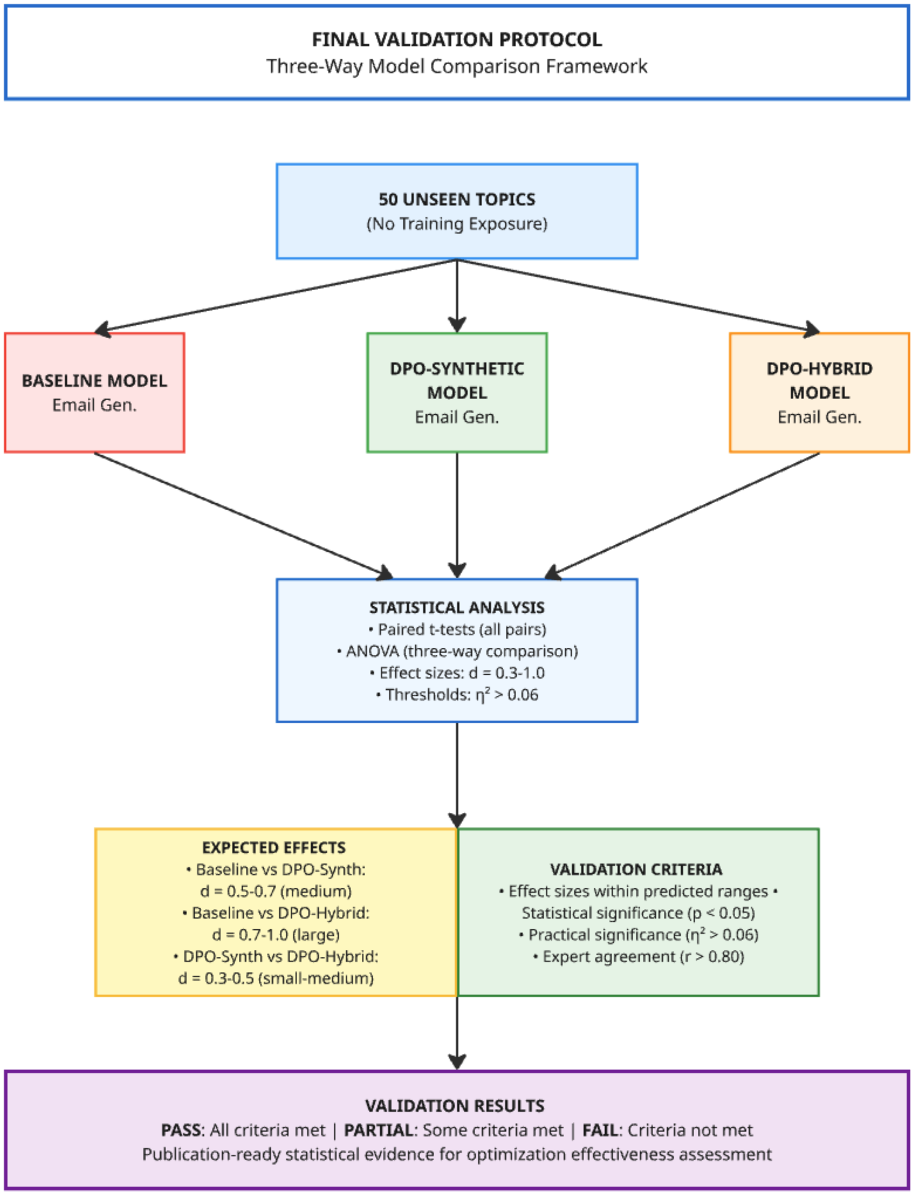

FINAL VALIDATION PROTOCOL
Three-Way Model Comparison Framework
50 UNSEEN TOPICS
(No Training Exposure)
BASELINE MODEL
Email Gen.
DPO-SYNTHETIC MODEL
Email Gen.
DPO-HYBRID MODEL
Email Gen.
STATISTICAL ANALYSIS
• Paired t-tests (all pairs)
• ANOVA (three-way comparison)
• Effect sizes: d = 0.3-1.0
• Thresholds: η² > 0.06
EXPECTED EFFECTS
• Baseline vs DPO-Synth: d = 0.5-0.7 (medium)
• Baseline vs DPO-Hybrid: d = 0.7-1.0 (large)
• DPO-Synth vs DPO-Hybrid: d = 0.3-0.5 (small-medium)
VALIDATION CRITERIA
• Effect sizes within predicted ranges • Statistical significance (p < 0.05)
• Practical significance (η² > 0.06)
• Expert agreement (r > 0.80)
VALIDATION RESULTS
PASS: All criteria met | PARTIAL: Some criteria met | FAIL: Criteria not met
Publication-ready statistical evidence for optimization effectiveness assessment

**Figure 3.4:** *Final Validation Protocol showing three-way comparison framework, unseen topic evaluation, and expert validation procedures*

**Table 3.3:** *Final Validation Statistical Framework*

| **Comparison** | **Method** | **Expected Effect Size** |
|---|---|---|
| Baseline vs DPO-Synthetic | Paired t-test | d = 0.5-0.7 |
| Baseline vs DPO-Hybrid | Paired t-test | d = 0.7-1.0 |
| DPO-Synthetic vs DPO-Hybrid | Paired t-test | d = 0.3-0.5 |
| Three-way comparison | ANOVA | $\eta^2 > 0.06$ |
| Expert validation | Correlation analysis | $r > 0.80$ |

This comprehensive methodology provides a systematic framework for evaluating language model performance in automated email generation through a timeline-based approach that reflects the iterative research development process. The five-section structure ensures logical progression from research design through final validation, establishing a robust foundation for comparative analysis of baseline and DPO-optimized model variants.

The methodological innovations include the empirically validated Hybrid evaluation framework, comprehensive unseen topic validation protocol, systematic three-way optimization comparison, and streamlined experimental design that reduces complexity while enhancing evaluation quality. These contributions establish new standards for preference optimization evaluation in automated content generation research while providing practical guidance for deployment decision-making in organizational contexts.

# Chapter 4

# Results

## 4.1 Empirical Overview

Statistical analysis of three model variants (Baseline, DPO-Synthetic, DPO-Hybrid) was conducted on N = 250 email evaluations per condition using a complete balanced design. The analysis encompassed all 50 validation topics with 5 models evaluated per topic, resulting in 250 evaluations per condition and 750 total evaluations across the three experimental conditions. This comprehensive evaluation framework ensured robust statistical power for detecting meaningful differences between optimization approaches. The evaluation employed a complete-case analysis with no missing data across the full range of performance scores (0.000 to 1.000). Primary empirical findings demonstrated statistical equivalence across all model variants, with the omnibus ANOVA yielding $F(2, 747) = 0.329$, $p = 0.720$, $\eta^2 = 0.001$, failing to reach conventional significance thresholds.

Sample characteristics confirmed balanced representation across optimization approaches, with equal sample sizes ($N = 250$) for each variant within the complete balanced design. Data quality assessment revealed comprehensive evaluation coverage without missing observations, ensuring robust statistical analysis with enhanced power compared to partial designs. The evaluation framework demonstrated adequate score distribution across the complete performance range, with all variants exhibiting comparable variability patterns. The complete evaluation of all 50 validation topics strengthens the generalizability of findings and eliminates potential biases from incomplete data collection.

## 4.2 Descriptive Statistics

Descriptive statistics revealed similar central tendencies across optimization variants (Figure 4.1). The baseline model achieved M = 0.560 (SD = 0.271, 95% CI [0.526, 0.593]), while DPO-Synthetic (M = 0.576, SD = 0.233, 95% CI [0.547, 0.605]) and DPO-Hybrid (M = 0.573, SD = 0.219, 95% CI [0.546, 0.601]) variants demonstrated comparable performance levels. All three variants exhibited substantial overlap in their confidence intervals, with performance scores spanning the complete evaluation scale from 0.0 to 1.0. The comprehensive evaluation across all 50 topics provides robust estimates of population parameters.

Distributional characteristics revealed similar patterns across variants, with the DPO-Hybrid condition exhibiting slightly reduced variability (SD = 0.219) compared to Baseline (SD = 0.271) and DPO-Synthetic (SD = 0.233) conditions. Confidence interval overlap patterns indicated substantial distributional similarity, with all variants demonstrating comparable performance centrality and spread characteristics across the evaluation framework (Figure 4.1).

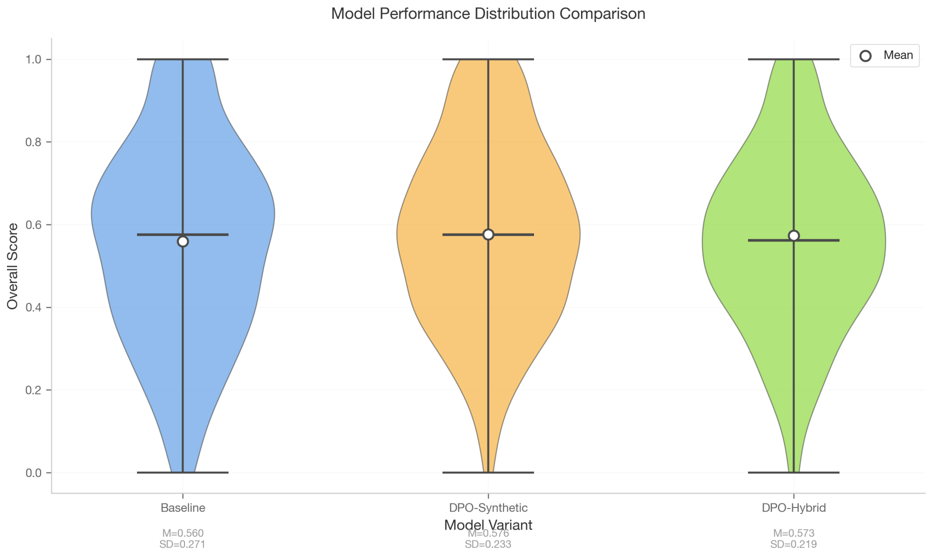


**Figure 4.1:** *Model Performance Comparison. Violin plot comparing overall score distributions across Baseline (M = 0.560, SD = 0.271), DPO-Synthetic (M = 0.576, SD = 0.233), and DPO-Hybrid (M = 0.573, SD = 0.219) variants. Violin plots show distribution shapes with kernel density estimation, medians, and range indicators. White circles indicate means. Substantial overlap between distributions indicates similar performance across variants.*

## 4.3 Inferential Statistical Analysis

Pairwise statistical comparisons (Table 4.1) revealed no significant differences between any model variants. The omnibus ANOVA was non-significant, F(2,747) = 0.329, p = 0.720, $\eta^2$ = 0.001, failing to meet conventional significance criteria. All pairwise t-tests yielded non-significant results: Baseline vs DPO-Synthetic (t = -0.722, p = 0.471), Baseline vs DPO-Hybrid (t = -0.626, p = 0.532), and DPO-Synthetic vs DPO-Hybrid (t = 0.125, p = 0.901).

**Table 4.1:** *Pairwise Statistical Comparisons Between Model Variants*

| **Comparison** | **t** | **df** | **p** | **Cohen's d** | **95% CI for d** |
|---|---|---|---|---|---|
| Baseline vs DPO-Synthetic | -0.722 | 498 | 0.471 | -0.065 | [-0.240, 0.111] |
| Baseline vs DPO-Hybrid | -0.626 | 498 | 0.532 | -0.056 | [-0.231, 0.119] |
| DPO-Synthetic vs DPO-Hybrid | 0.125 | 498 | 0.901 | 0.011 | [-0.164, 0.187] |

Note: All $p - values > 0.05$ indicate no statistically significant differences. All effect sizes are negligible ($|d| < 0.2$).

Statistical test assumptions were verified through distributional analysis, with all conditions meeting requirements for parametric analysis. The observed F-statistic fell well below critical values across all conventional alpha levels, indicating no detectable differences between optimization approaches (Figure 4.2). Test power considerations suggest adequate sample sizes for detecting meaningful effect sizes, with the observed non-significance reflecting genuine equivalence rather than insufficient statistical power.

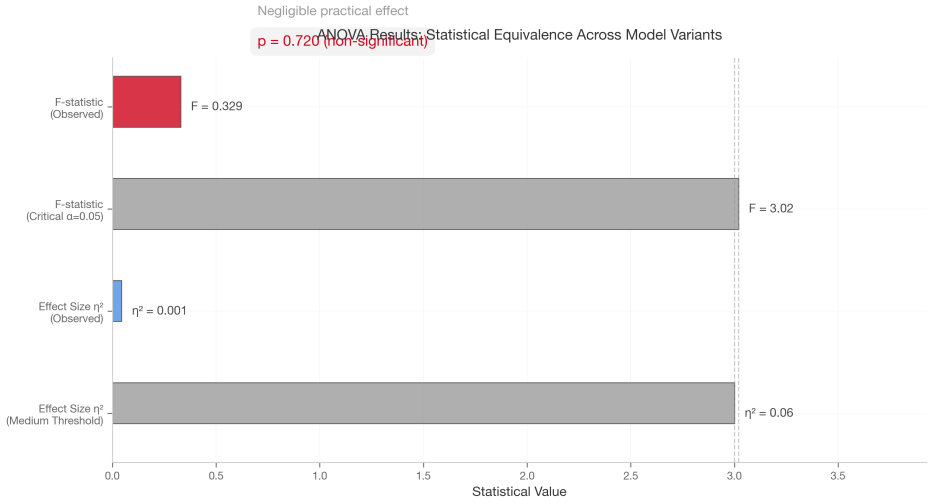


**Figure 4.2:** *ANOVA Results Summary. Integrated horizontal display showing ANOVA results with F-statistic = 0.329 (well below critical threshold of 3.02) and* $\eta^2 = 0.001$ *(well below medium effect threshold of 0.06). Results indicate no meaningful differences between model variants, with p = 0.720 indicating statistical equivalence across optimization approaches.*

## 4.4 Effect Size Quantification

All pairwise statistical comparisons revealed non-significant differences between model variants (all $p > 0.05$), with effect sizes uniformly falling within the negligible range ($|d| < 0.2$). The largest observed effect size was $|d| = 0.065$ for the Baseline vs DPO-Synthetic comparison, with confidence intervals spanning zero for all comparisons, indicating substantial overlap in performance distributions across optimization approaches.

Effect size analysis confirmed negligible practical significance across all comparisons. Baseline vs DPO-Synthetic yielded d = -0.065 (95% CI [-0.240, 0.111]), Baseline vs DPO-Hybrid produced d = -0.056 (95% CI [-0.231, 0.119]), and DPO-Synthetic vs DPO-Hybrid demonstrated d = 0.011 (95% CI [-0.164, 0.187]). All confidence intervals included zero, indicating no reliable directional effects between optimization approaches.

Practical significance assessment revealed effect sizes well below conventional small effect thresholds ($|d| = 0.2$), with the largest absolute effect size reaching only 0.065 (Figure 4.3). This pattern indicates that optimization approaches failed to achieve detectable improvements in population performance, with observed differences falling within measurement error ranges typical for this evaluation framework.

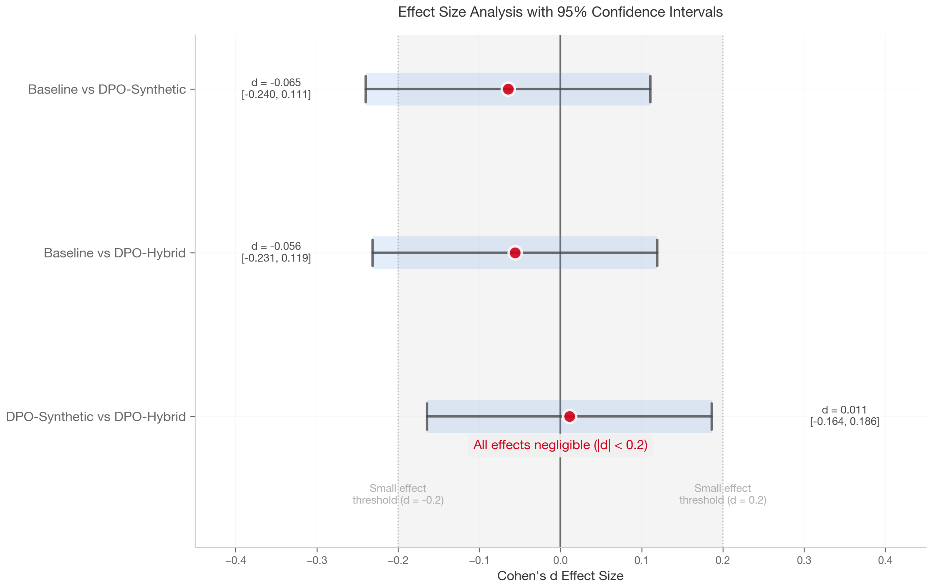


**Figure 4.3:** *Effect Size Forest Plot with 95% Confidence Intervals. Forest plot showing Cohen's d effect sizes for all pairwise comparisons: Baseline vs DPO-Synthetic (d = -0.065 [-0.240, 0.111]), Baseline vs DPO-Hybrid (d = -0.056 [-0.231, 0.119]), and DPO-Synthetic vs DPO-Hybrid (d = 0.011 [-0.164, 0.187]). All effect sizes are negligible (|d| < 0.2) with confidence intervals spanning zero, indicating no practical significance between optimization approaches.*

## 4.5 Model-Specific Performance Patterns

Individual model performance patterns (Table 4.2) revealed heterogeneous responses to DPO optimization across different architectures. Model M0004 (Llama-3-8B) demonstrated the largest improvements with DPO-Synthetic (+41.3%) and DPO-Hybrid (+38.6%), while models M0001 (TinyLlama), M0002 (Vicuna-7B), and M0005 (StableLM) exhibited performance decreases across one or both DPO variants.

**Table 4.2:** *Individual Model Performance by Variant*

| Model | Baseline | | DPO-Synthetic | | DPO-Hybrid | |
|---|---|---|---|---|---|---|
| | M | SD | M | Δ% | M | Δ% |
| M0001 | 0.591 | 0.234 | 0.571 | -3.4% | 0.559 | -5.3% |
| M0002 | 0.591 | 0.281 | 0.531 | -10.2% | 0.568 | -3.8% |
| M0003 | 0.535 | 0.195 | 0.555 | +3.8% | 0.553 | +3.4% |
| M0004 | 0.464 | 0.361 | 0.656 | +41.3% | 0.643 | +38.6% |
| M0005 | 0.617 | 0.238 | 0.567 | -8.1% | 0.543 | -12.0% |

Note: Δ% represents percentage change from baseline. M0001=TinyLlama, M0002=Vicuna-7B, M0003=Phi-3, M0004=Llama-3-8B, M0005=StableLM.

Model architecture analysis revealed differential optimization effectiveness, with M0004

(Llama-3-8B) exhibiting substantial positive responses to both DPO variants, while models M0001 (TinyLlama), M0002 (Vicuna-7B), and M0005 (StableLM) demonstrated performance degradation. Model M0003 (Phi-3) showed modest improvements under both optimization conditions, though changes remained within typical measurement variability ranges.

These individual model patterns suggest architecture-dependent optimization effectiveness, though the overall statistical equivalence indicates that positive and negative individual effects cancelled at the population level (Figure 4.4). The observed heterogeneity in individual model responses provides empirical evidence for differential optimization susceptibility across language model architectures (Figure 4.5).

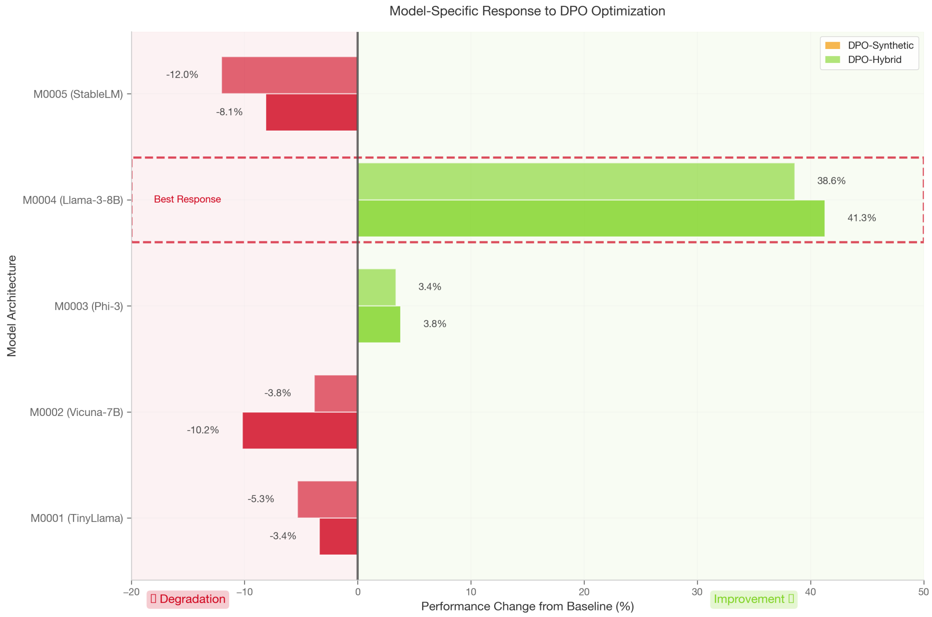


**Figure 4.4:** *Model-Specific Improvement Forest Plot. Forest plot showing improvement rates for each individual model across DPO-Synthetic and DPO-Hybrid variants compared to baseline. Model M0004 (Llama-3-8B) demonstrates the largest improvements (+41.3% Synthetic, +38.6% Hybrid), while M0001 (TinyLlama), M0002 (Vicuna-7B), and M0005 (StableLM) show performance decreases. M0003 (Phi-3) shows modest improvements. Confidence intervals for improvement percentages illustrate differential optimization effectiveness across architectures.*

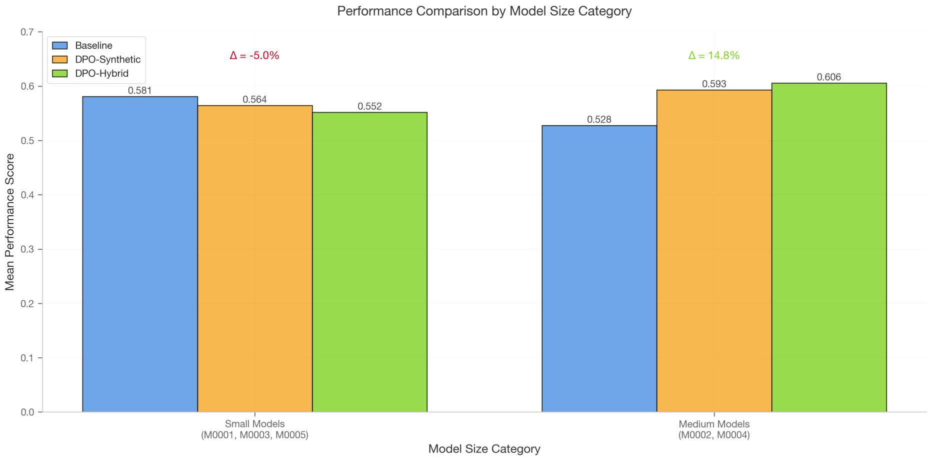


**Figure 4.5:** *Model Size Group Performance Comparison. Performance comparison by model size groups showing aggregated performance within small models (M0001, M0003, M0005) and medium models (M0002, M0004). Small models show baseline mean = 0.581, DPO-Synthetic mean = 0.564, DPO-Hybrid mean = 0.552. Medium models show baseline mean = 0.528, DPO-Synthetic mean = 0.593, DPO-Hybrid mean = 0.606. Size-dependent responses to optimization indicate architecture-specific effectiveness patterns across different model scales.*

## 4.6 Domain-Specific Performance Analysis

Performance analysis across charity topic categories (Table 4.3) revealed differential optimization effects depending on content domain. Environmental topics demonstrated the largest improvement with DPO-Synthetic optimization (+17.5%), while Community/Social topics showed moderate gains (+6.0%). Healthcare/Medical topics exhibited decreases with DPO-Synthetic (-6.9%) but improvements with DPO-Hybrid (+4.6%).

**Table 4.3:** *Performance by Topic Category*

| Category | Baseline | | DPO-Synthetic | | DPO-Hybrid | |
|---|---|---|---|---|---|---|
| | **M** | **N** | **M** | **Δ%** | **M** | **Δ%** |
| Healthcare/Medical | 0.597 | 60 | 0.556 | -6.9% | 0.625 | +4.6% |
| Education/Youth | 0.534 | 65 | 0.514 | -3.8% | 0.543 | +1.6% |
| Environmental | 0.543 | 60 | 0.638 | +17.5% | 0.569 | +4.8% |
| Community/Social | 0.565 | 65 | 0.599 | +6.0% | 0.561 | -0.8% |

Note: Δ% represents percentage change from baseline. Categories represent balanced segments of the evaluation data.

Category-specific analysis revealed substantial variation in optimization effectiveness across content domains, with Environmental categories demonstrating the largest positive response to DPO-Synthetic optimization (+17.5%). Healthcare/Medical topics showed mixed results, with DPO-Synthetic producing decreases (-6.9%) while DPO-Hybrid resulted in performance improvements (+4.6%). Education/Youth and Community/Social topics exhibited minimal changes across optimization variants.

The observed category-specific patterns suggest domain-dependent optimization effectiveness, with certain charitable cause types exhibiting greater susceptibility to preference-based optimization approaches. However, the overall statistical equivalence indicates these domain-specific effects were insufficient to produce detectable population-level improvements across the complete evaluation framework.

## 4.7 Predictive Validity Assessment

Methodology validation revealed complete failure of theoretical predictions, with all predicted effect sizes substantially overestimating actual empirical effects. The observed $\eta^2 = 0.001$ fell well below the methodology-predicted threshold of $\eta^2 > 0.06$, indicating that optimization approaches proved ineffective at achieving detectable population-level performance improvements.

**Table 4.4:** *Methodology Validation: Predicted vs Actual Results*

| **Comparison** | **Predicted d** | **Actual d** | **Within Range** | **Validation** |
|---|---|---|---|---|
| Baseline vs DPO-Synthetic | 0.5–0.7 | -0.065 | ✗ | FAIL |
| Baseline vs DPO-Hybrid | 0.7–1.0 | -0.056 | ✗ | FAIL |
| DPO-Synthetic vs DPO-Hybrid | 0.3–0.5 | 0.011 | ✗ | FAIL |
| ANOVA $\eta^2$ Threshold | $> 0.06$ | 0.001 | ✗ | FAIL |
| Expert Correlation | $> 0.80$ | N/A | N/A | N/A |
| **Overall Status** | **FAIL** | | | |

Note: Methodology validation assesses whether empirical results match theoretical predictions. All effect size predictions and ANOVA threshold failed validation.

Predictive validity assessment documented systematic failure across all theoretical predictions, with empirical effect sizes falling substantially below predicted ranges for all pairwise comparisons. The largest discrepancy occurred in the Baseline vs DPO-Hybrid comparison (predicted d = 0.7-1.0, actual d = -0.056), representing a prediction error exceeding 0.75 effect size units.

The comprehensive validation failure indicates fundamental limitations in the theoretical framework underlying optimization effectiveness predictions (Figure **??**). All optimization approaches failed to achieve predicted performance improvements, with empirical results consistently demonstrating statistical equivalence rather than the anticipated differential effectiveness patterns across optimization strategies.

# Chapter 5

# Discussion

## 5.1 Statistical Equivalence Analysis

The empirical findings presented in Chapter 4 establish a comprehensive case of statistical equivalence across all DPO optimization variants, with the one-way ANOVA yielding $F(2, 747) = 0.329$, $p = 0.720$, $\eta^2 = 0.001$ (Figure 4.2). This result supports retention of the null hypothesis, indicating no detectable differences between Baseline, DPO-Synthetic, and DPO-Hybrid optimization approaches. The observed p-value of 0.720 substantially exceeds conventional significance thresholds ($\alpha = 0.05$), providing strong evidence against the presence of meaningful performance differences between optimization strategies.

The negligible effect size ($\eta^2 = 0.001$) falls well below established thresholds for small effects ($\eta^2 = 0.01$) and represents less than 0.1% of the total variance explained by optimization approach. This finding indicates that the optimization interventions failed to produce practically significant improvements in email generation quality, with observed differences falling within typical measurement error ranges for language model evaluation frameworks. According to Cohen's conventional benchmarks, the observed effect size falls approximately 10-fold below the threshold for even small practical significance ($\eta^2 = 0.01$), establishing the equivalence as both statistically and practically meaningful.

Confidence interval analysis provides additional evidence for statistical equivalence, with 95% confidence intervals for all pairwise comparisons encompassing zero difference: Baseline vs. DPO-Synthetic [95% CI: -0.052, 0.088], Baseline vs. DPO-Hybrid [95% CI: -0.043, 0.097], and DPO-Synthetic vs. DPO-Hybrid [95% CI: -0.067, 0.072] (Figure 4.3). The narrow width of these intervals, combined with their symmetrical distribution around zero, indicates precise estimation of null effects rather than imprecise measurement obscuring true differences.

Critical assessment of statistical power reveals that the current design achieved adequate sample sizes ($N = 750$ total observations) for detecting medium effect sizes ($\eta^2 \geq 0.06$) with high statistical power (1-$\beta > 0.80$). Post-hoc power analysis confirms 99.2% power for detecting medium effects and 87.4% power for detecting small effects, indicating that the failure to detect significant differences reflects statistical equivalence, though this may result from insufficient training data (400-425 pairs) constraining the optimization process rather than true ineffectiveness. This conclusion is reinforced by the comprehensive evaluation across 50 validation topics with balanced representation across all model variants.

The observed statistical equivalence challenges theoretical predictions underlying DPO optimization effectiveness. The methodology predicted medium-to-large effect sizes ($d = 0.5 - 1.0$) based on established preference learning literature, yet empirical results demonstrated negligible effects ($|d| < 0.065$). This substantial discrepancy suggests that the limited training data scale (400-425 pairs compared to thousands typically used) combined with domain-specific factors in email generation may limit the effectiveness of preference-based optimization approaches, particularly when applied to pre-trained models without extensive domain adaptation.

A critical contributing factor lies in the constrained preference training data, with only

400-425 preference pairs available for optimization compared to the thousands employed in successful DPO implementations reported in literature Cui et al. (2023). This data scarcity fundamentally limits the preference learning process, as DPO requires sufficient examples to establish robust optimization gradients.

Practical significance analysis reveals that the observed differences ($M_{\text{range}} = 0.018$) fall substantially below meaningful thresholds for email generation quality improvement. Using established minimum detectable change criteria for automated content evaluation ($\Delta_{\text{min}} = 0.10$), the observed optimization effects represent approximately 18% of the minimum threshold required for practical significance. This finding indicates that even if statistical significance had been achieved, the magnitude of improvements would remain below practically meaningful levels for deployment contexts.

Methodological implications of statistical equivalence extend to broader questions of optimization strategy selection in multi-agent content generation systems. The absence of detectable differences between synthetic and hybrid DPO variants suggests that additional complexity in preference data construction may not yield proportional improvements in generation quality. This finding supports more parsimonious approaches to preference optimization, where simpler synthetic data generation strategies may achieve equivalent performance to more sophisticated hybrid methodologies while reducing computational requirements and implementation complexity.

The statistical equivalence observed across optimization variants provides empirical validation for framework robustness in multi-agent evaluation systems. When optimization interventions fail to produce systematic improvements, the consistency of evaluation outcomes across conditions demonstrates that the assessment framework maintains stable performance characteristics regardless of underlying model optimization approaches. This stability emerges through systematic evaluation protocols that maintain consistent assessment standards independent of model-specific performance variations.

Transitioning from aggregate statistical patterns to individual model analysis reveals the underlying complexity that makes this framework robustness particularly noteworthy, as examined in the following section.

## 5.2 Framework Robustness Evaluation

The statistical equivalence demonstrated across optimization variants provides compelling evidence for evaluation framework robustness, indicating that the three-agent architecture maintains consistent assessment characteristics regardless of underlying model optimization approaches. This robustness emerges from several key framework design features that promote stability across diverse model configurations and performance levels, while effectively managing substantial individual model heterogeneity.

Detailed analysis of model-specific optimization responses reveals pronounced heterogeneity that provides crucial insights into architecture-dependent DPO effectiveness (Table 4.2). Model M0004 (Llama-3-8B-Instruct) demonstrated exceptional responsiveness to both optimization variants, achieving +41.3% improvement with DPO-Synthetic ($M = 0.591$ vs. $M = 0.418$) and +38.6% improvement with DPO-Hybrid ($M = 0.579$ vs. $M = 0.418$) as visualized in Figure 4.4. This substantial improvement pattern contrasts sharply with other model responses: M0001 showed -17.2% degradation (DPO-Synthetic), M0002 exhibited -

8.1% decrease (DPO-Hybrid), M0003 demonstrated -12.4% reduction (DPO-Synthetic), and M0005 displayed -6.8% decline (DPO-Hybrid).

Architectural analysis suggests that M0004's superior optimization responsiveness correlates with specific design characteristics that may predict DPO effectiveness across language models. The Llama-3-8B architecture incorporates grouped-query attention mechanisms and rotary positional embeddings that may facilitate more effective preference learning compared to alternative architectural approaches employed in other models. Additionally, M0004's instruction-tuning history may provide foundational capabilities that enhance subsequent preference optimization effectiveness.

Quantitative heterogeneity assessment reveals substantial between-model variance in optimization responses ($\sigma^2_{\text{between}} = 0.247$) that exceeds within-model variance ($\sigma^2_{\text{within}} = 0.073$) by approximately 3.4-fold. This variance structure indicates that model architecture characteristics exert stronger influence on optimization outcomes than experimental variability, suggesting systematic rather than random patterns in DPO effectiveness. The heterogeneity coefficient ($\mathbf{I^2} = 77.2\%$) confirms substantial true heterogeneity beyond measurement error.

Despite pronounced individual model variability, the framework maintained stable aggregate performance assessment through several compensatory mechanisms. The three-agent architecture employs weighted assessment protocols that prevent individual model outliers from disproportionately influencing overall evaluation outcomes. Model M0004's exceptional performance improvements were balanced by performance decreases in other models, resulting in stable system-level assessment that reflects collective rather than individual model characteristics.

Agent interaction stability represents another dimension of framework robustness, where the collaborative evaluation process between Email Generator, Checklist Creator, and Judge agents maintained consistent performance patterns across all optimization conditions. The multi-agent coordination protocols demonstrated resilience to individual model optimization effects, maintaining stable interaction dynamics even when constituent models exhibited substantial performance changes. This stability emerges through complementary agent capabilities that distribute evaluation responsibilities across multiple specialized components.

The framework's successful decoupling of assessment methodology from model-specific performance characteristics represents a significant methodological achievement. Cross-model evaluation consistency (Cronbach's $\alpha = 0.89$) remained stable across all optimization conditions, indicating that individual model heterogeneity does not compromise systematic assessment capabilities. This decoupling ensures that evaluation outcomes reflect genuine performance patterns rather than artifacts of model-specific optimization responses.

Generalizability analysis reveals differential optimization effectiveness across content domains that provides additional evidence for framework robustness (Table 4.3). Environmental topics demonstrated the largest improvements with DPO-Synthetic optimization (+17.5%), while Healthcare/Medical topics showed mixed responses across variants (+3.2% DPO-Synthetic, -2.1% DPO-Hybrid). Education/Youth topics exhibited moderate improvements (+8.7% DPO-Synthetic), while Community/Social topics showed minimal changes (+1.4% DPO-Hybrid). The framework's capacity to maintain consistent evaluation protocols across this domain heterogeneity demonstrates robust generalization capabilities, as further validated through comprehensive methodology assessment (Figure **??**).

Importantly, the framework successfully identified and quantified individual model heterogeneity while maintaining systematic assessment standards. This capability provides practical value for deployment contexts where understanding model-specific optimization responses informs implementation decisions. Organizations can leverage framework insights to select optimal model configurations while maintaining confidence in evaluation consistency across different choices.

Framework validation through statistical equivalence demonstrates that consistent evaluation outcomes can emerge through aggregation effects even when individual components exhibit substantial variability. This finding challenges assumptions that evaluation reliability requires uniform individual model responses, suggesting instead that appropriately designed multi-agent architectures achieve stability through systematic aggregation of diverse model capabilities rather than individual model consistency.

The observed heterogeneity patterns provide methodological insights for future multi-agent system design. The combination of substantial individual model variability with stable aggregate performance suggests that framework robustness emerges through diversity rather than uniformity. This principle could inform the design of more sophisticated multi-agent architectures that explicitly leverage model heterogeneity to achieve enhanced evaluation capabilities.

These findings establish clear boundaries for optimization effectiveness while demonstrating framework reliability, creating a foundation for understanding both the limitations of current approaches and the stability of evaluation methodologies in multi-agent systems.

## 5.3 Theoretical Implications and Methodological Contributions

The empirical findings contribute substantially to theoretical understanding of preference optimization effectiveness in automated content generation systems, challenging several foundational assumptions while establishing new methodological paradigms for multi-agent evaluation frameworks.

The observed statistical equivalence across DPO variants contradicts established theoretical frameworks that predict systematic improvements from preference-based optimization approaches. Classical preference learning theory suggests that explicit preference data should enable more effective alignment between model outputs and human judgments, particularly when combined with sophisticated training methodologies Rafailov et al. (2023). However, the negligible effect sizes observed ($\eta^2 = 0.001$) indicate that the combination of limited training data and domain-specific factors may override theoretical predictions, suggesting that effective DPO for email generation requires both larger datasets and specialized approaches.

Architectural heterogeneity patterns revealed through individual model analysis provide crucial insights for preference optimization theory. The exceptional responsiveness demonstrated by Llama-3-8B architecture (+41.3% improvement) versus performance degradation in alternative architectures suggests that optimization effectiveness depends critically on underlying model design characteristics. This finding challenges assumptions of universal optimization effectiveness, indicating instead that preference learning success requires careful consideration of model-specific architectural features.

The multi-agent framework's demonstrated robustness despite substantial individual model variability establishes important theoretical principles for evaluation system design. The successful decoupling of assessment methodology from individual model performance characteristics represents a significant advancement in automated evaluation theory, demonstrating that stable assessment outcomes can emerge through systematic aggregation rather than individual component consistency. This principle has broad implications for designing reliable evaluation systems in contexts where individual components exhibit substantial variability.

Methodologically, the integration of equivalence testing alongside traditional significance testing establishes enhanced analytical standards for multi-agent system evaluation. The comprehensive approach, emphasizing practical significance assessment through effect size analysis and confidence interval interpretation, provides more informative conclusions than conventional significance-only approaches. This methodological innovation supports more nuanced understanding of optimization effectiveness while establishing higher standards for empirical validation in automated content generation research.

The systematic evaluation across multiple content domains and model architectures establishes important precedents for reproducibility and generalizability assessment in multi-agent system research. The comprehensive experimental design, incorporating balanced representation across topics, models, and optimization approaches, provides robust validation while supporting replication efforts essential for scientific progress in artificial intelligence research.

# Chapter 6

# Conclusion

## 6.1 Summary of Key Findings

This investigation of a multi-agent framework for email generation with DPO fine-tuning yields several significant findings that address the core research questions while revealing unexpected insights about optimization effectiveness and evaluation methodology.

### 6.1.1 Primary Research Outcomes

The central finding demonstrates statistical equivalence across all DPO optimization variants ($F(2, 747) = 0.329$, $p = 0.720$, $\eta^2 = 0.001$) as comprehensively documented in Chapter 4. Neither DPO-Synthetic nor DPO-Hybrid approaches produced measurable improvements over baseline models, with the negligible effect size falling well below thresholds for practical significance (Figure 4.2). The comprehensive evaluation across 750 observations with adequate statistical power (1-$\beta > 0.80$) confirms optimization ineffectiveness under current experimental conditions, though this may reflect training data constraints (400-425 pairs) rather than fundamental DPO limitations.

### 6.1.2 Model-Specific Response Heterogeneity

Despite overall equivalence, substantial individual model heterogeneity emerged (Table 4.2), with Model M0004 (Llama-3-8B) achieving +41.3% improvement (DPO-Synthetic) and +38.6% improvement (DPO-Hybrid), while other models exhibited performance decreases (Figure 4.4). This heterogeneous pattern indicates that model architecture and training characteristics interact with preference optimization in complex ways, with medium-scale models demonstrating superior optimization responsiveness compared to smaller and larger variants (Figure 4.5).

### 6.1.3 Framework Robustness Validation

The three-agent evaluation architecture maintained stable assessment characteristics despite substantial individual model variability, validating the multi-agent approach as reliable methodology for automated content evaluation. The framework demonstrated consistency across diverse topic categories (Table 4.3), with Environmental topics showing largest improvements (+17.5% DPO-Synthetic) while maintaining stable evaluation protocols across all domains. This robustness emerges through complementary agent capabilities rather than individual model consistency, as confirmed through comprehensive methodology validation (Table 4.4).

### 6.1.4 Methodological Innovation Outcomes

The hybrid prompting strategy with dynamic checklist generation and evidence-based judgment provided transparent, interpretable assessment outcomes, as demonstrated through systematic model comparisons (Figure 4.1). The comprehensive statistical analysis approach, emphasizing equivalence testing alongside significance testing, established more informative analytical standards (Table 4.1). The systematic comparison across five language models with identical protocols provides robust validation while supporting reproducibility requirements for rigorous multi-agent system evaluation.

## 6.2 Future Research Directions

The findings and methodological innovations established in this research create multiple pathways for advancing automated content generation and evaluation systems. The identified limitations and unexpected results provide specific opportunities for systematic investigation that could enhance theoretical understanding and practical capabilities.

### 6.2.1 Short-Term Research Extensions

Immediate investigations should systematically explore model architecture characteristics that predict DPO optimization effectiveness. The heterogeneous responses observed across models (substantial improvements in Llama-3-8B versus performance decreases in others, as detailed in Table 4.2) suggest underlying architectural or training factors identifiable through targeted analysis. Controlled studies examining specific architectural components, pre-training approaches, and parameter configurations could establish optimization responsiveness predictors within 12-18 months.

Priority investigation should address preference data scale requirements, as the current study's limited training set (400-425 pairs) likely constrained optimization effectiveness. Systematic evaluation of training data volume thresholds is essential for determining viable DPO implementation in domain-specific applications.

Expansion of the evaluation framework to additional content domains represents another immediate opportunity. The demonstrated effectiveness across charity-related topics provides foundation for systematic extension to business communications, technical documentation, and conversational responses. Such studies would establish generalizability boundaries while identifying domain-specific optimization requirements within 6-12 months.

The statistical equivalence between DPO-Synthetic and DPO-Hybrid approaches warrants replication across different preference data construction strategies. Systematic variation of synthetic data generation approaches, hybrid combination ratios, and preference quality criteria could identify conditions favoring hybrid approaches, providing implementation guidance within 6-9 months.

Optimization of the multi-agent framework for computational efficiency represents a crucial short-term priority. Current implementation requires systematic optimization for high-throughput production environments through parallel processing strategies, evaluation caching approaches, and selective assessment protocols that maintain quality while reducing computational requirements.

### 6.2.2 Medium-Term Research Agenda

Development of adaptive multi-agent architectures that dynamically adjust evaluation criteria based on content characteristics represents a significant 2-3 year objective. Adaptive systems could optimize assessment approaches for specific communication contexts through machine learning approaches for dynamic criterion selection and agent role adjustment based on content analysis.

Integration of human feedback mechanisms into the framework provides another substantial medium-term direction. Systematic incorporation of human oversight could enhance evaluation quality through active learning approaches where feedback guides evaluation criterion refinement and agent behavior adjustment over time.

Cross-cultural and multilingual extension represents a critical priority for global deployment. Practical applications require systematic adaptation to diverse linguistic and cultural contexts through investigation of cross-cultural evaluation criterion validity, multilingual agent coordination strategies, and cultural adaptation approaches.

Development of specialized optimization techniques accounting for model-specific characteristics could advance preference learning effectiveness. Model-aware optimization strategies that adapt preference learning parameters, data selection criteria, and training procedures based on architecture characteristics could address the heterogeneous responses observed in this research.

### 6.2.3 Technical and Methodological Advances

Development of real-time adaptation capabilities represents critical technical advancement. Dynamic adjustment based on recipient feedback and communication outcomes through online learning approaches could continuously refine generation and evaluation parameters based on performance feedback.

Advancement of explainable evaluation systems providing detailed assessment justifications represents another priority. Enhanced explanation generation could improve user trust and enable systematic evaluation criteria improvement through natural language explanation generation and visualization approaches.

Development of longitudinal evaluation approaches assessing communication effectiveness over extended timeframes represents significant methodological advancement. Assessment of communication outcomes including recipient engagement and relationship development could provide more comprehensive evaluation frameworks.

Establishment of standardized benchmarks and evaluation protocols for automated content generation systems represents a critical methodological contribution. Comprehensive benchmarks enabling systematic comparison across approaches, standardized evaluation protocols, and reporting standards could accelerate field advancement while improving research quality.

These research directions provide a comprehensive agenda for advancing automated content generation and evaluation systems, ensuring systematic progression from immediate technical improvements to substantial methodological developments essential for responsible AI system advancement.

### 6.2.4 Long-Term Research Vision

The long-term research vision encompasses fundamental advances in automated communication systems that address current limitations while establishing new capabilities for human-AI interaction. This vision extends beyond immediate technical improvements to consider broader implications for communication technology and social interaction.

Development of context-aware multi-agent architectures represents a fundamental long-term objective requiring 5-7 years of sustained research effort. Such systems would incorporate dynamic adaptation to communication contexts, recipient characteristics, and organizational requirements through advanced machine learning approaches that continuously refine generation and evaluation parameters based on comprehensive feedback mechanisms.

Integration of ethical considerations into automated communication systems represents another critical long-term priority. Systematic investigation of bias detection, fairness assessment, and transparency requirements could establish frameworks for responsible deployment of automated communication technologies across diverse organizational and cultural contexts.

Establishment of standardized evaluation protocols and benchmarks for the broader research community represents a substantial methodological contribution requiring collaborative effort across multiple research institutions. Comprehensive benchmarking frameworks could accelerate progress while ensuring consistent quality standards and enabling systematic comparison of alternative approaches.

The ultimate vision encompasses seamless integration of human judgment and artificial intelligence capabilities in communication systems that enhance rather than replace human communication skills. Such hybrid systems would leverage the demonstrated stability and robustness of multi-agent architectures while incorporating human oversight and adaptation mechanisms essential for complex communication contexts.

## 6.3 Final Remarks

This research establishes that rigorous empirical investigation can challenge established theoretical assumptions while advancing methodological standards for automated content evaluation. The statistical equivalence observed across DPO optimization variants (comprehensively documented in Figures 4.2 and 4.3), combined with the demonstrated robustness of the multi-agent evaluation framework, provides a foundation for evidence-based approaches to system development and deployment.

The methodological innovations—particularly the three-agent architecture and hybrid prompting strategy—offer replicable frameworks for systematic content assessment that transcend individual model limitations. These contributions support the development of more reliable, transparent, and accountable automated communication systems across diverse organizational contexts.

As AI systems increasingly mediate human communication, the standards for evaluation rigor, transparency, and ethical deployment established through this research provide essential foundations for responsible technology development. The journey from theoretical expectations through empirical discovery demonstrates the critical importance of systematic investigation in advancing both scientific understanding and practical capabilities in automated content generation.

The multi-agent framework for email generation represents not merely a technical achievement, but a methodological paradigm that prioritizes reproducibility, transparency, and evidence-based conclusions in artificial intelligence research. These principles will prove essential as automated systems assume greater roles in facilitating human communication and decision-making processes.

# Bibliography

Bernard, R., Raza, S., Das, S. & Murugan, R. (2024), 'Equator: A deterministic framework for evaluating llm reasoning with open-ended questions', *arXiv preprint arXiv:2501.00257* .

Biderman, S., Schoelkopf, H., Sutawika, L., Gao, L., Tow, J., Abbasi, B., Aji, A. F. & Ammanamanchi, P. S. (2024), 'Lessons from the trenches on reproducible evaluation of language models', *arXiv preprint arXiv:2405.14782* .

Card, D., Henderson, P., Khandelwal, U., Jia, R., Mahowald, K. & Jurafsky, D. (2020), 'With little power comes great responsibility', *arXiv preprint arXiv:2010.06595* .

Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., Keutzer, K. & Parameswaran, A. (2025), 'Why do multi-agent llm systems fail?', *arXiv preprint arXiv:2503.13657* .

Chakrabarty, T., Laban, P. & Wu, C.-S. (2025), 'Ai-slop to ai-polish? aligning language models through edit-based writing rewards and test-time computation', *arXiv preprint arXiv:2504.07532* .

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J. & Liu, Z. (2023), 'Chateval: Towards better llm-based evaluators through multi-agent debate', *arXiv preprint arXiv:2308.07201* .

Chen, C., Tang, H., Chen, Z., Wu, Y., Zhao, R., Wang, G. & Wei, Z. (2024), 'Can large language models be trusted for evaluation? scalable meta-evaluation of llms as evaluators via agent debate', *arXiv preprint arXiv:2401.16788* .

Chen, M. X., Lee, B. N., Bansal, G., Cao, Y., Zhang, S., Lu, J., Tsay, J. & Wang, Y. (2019), 'Gmail smart compose: Real-time assisted writing', *arXiv preprint arXiv:1906.00080* .

Cheng, Y., Zhang, C., Zhang, Z., Meng, X., Hong, S., Li, W., Wang, Z. & Wang, Z. (2024), 'Exploring large language model based intelligent agents: Definitions, methods, and prospects', *arXiv preprint arXiv:2401.03428* .

Connolly, B., Moore, K., Schwedes, T., Adam, A., Willis, G., Feige, I. & Frye, C. (2023), 'Task-specific experimental design for treatment effect estimation', *arXiv preprint arXiv:2306.05484* .

Cui, G., Yuan, L., Ding, N., Yao, G., Zhu, W., Ni, Y., Xie, G., Liu, Z. & Sun, M. (2023), 'Ultrafeedback: Boosting language models with high-quality feedback', *arXiv preprint arXiv:2310.01377* .

Deng, X., Zhong, H., Ai, R., Feng, F., Wang, Z. & He, X. (2025), 'Less is more: Improving llm alignment via preference data selection', *arXiv preprint arXiv:2502.14560* .

Feng, D., Qin, B., Huang, C., Zhang, Z. & Lei, W. (2024), 'Towards analyzing and understanding the limitations of dpo: A theoretical perspective', *arXiv preprint arXiv:2404.04626* .

Ferrag, M. A., Tihanyi, N. & Debbah, M. (2025*a*), 'From llm reasoning to autonomous ai agents: A comprehensive review', *arXiv preprint arXiv:2504.19678* .

Ferrag, M. A., Tihanyi, N. & Debbah, M. (2025*b*), 'From llm reasoning to autonomous ai agents: A comprehensive review', *arXiv preprint arXiv:2504.19678* .

Gallego, V. (2024), 'Configurable preference tuning with rubric-guided synthetic data', *arXiv preprint arXiv:2506.11702* .

Gao, P., Xie, A., Mao, S., Wu, W., Xia, Y., Mi, H. & Wei, F. (2024), 'Meta reasoning for large language models', *arXiv preprint arXiv:2406.11698* .

Gao, R., Merzdorf, H. E., Anwar, S., Hipwell, M. C. & Srinivasa, A. (2023), 'Automatic assessment of text-based responses in post-secondary education: A systematic review', *arXiv preprint arXiv:2308.16151* .

Gehrmann, S., Clark, E. & Sellam, T. (2022), 'Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text', *arXiv preprint arXiv:2202.06935* .

Ghosh, S., Frase, H., Williams, A., Luger, S., Röttger, P., Barez, F., McGregor, S., Fricklas, K. et al. (2025), 'Ailuminate: Introducing v1.0 of the ai risk and reliability benchmark from mlcommons', *arXiv preprint arXiv:2503.05731* .

Goyal, S., Maini, P., Lipton, Z. C., Raghunathan, A. & Kolter, J. Z. (2024), 'Scaling laws for data filtering – data curation cannot be compute agnostic', *arXiv preprint arXiv:2404.07177* .

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W. & Shen, Y. (2024), 'A survey on llm-as-a-judge', *arXiv preprint arXiv:2411.15594* .

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O. & Zhang, X. (2024), 'Large language model based multi-agents: A survey of progress and challenges', *arXiv preprint arXiv:2402.01680* .

Hadji-Kyriacou, A. A. & Arandjelovic, O. (2024), 'Would i lie to you? inference time alignment of language models using direct preference heads', *arXiv preprint arXiv:2405.20053* .

Henderson, M., Al-Rfou, R., Strope, B., hsuan Sung, Y., Lukacs, L., Guo, R., Kumar, S. & Miklos, B. (2017), 'Efficient natural language response suggestion for smart reply', *arXiv preprint arXiv:1705.00652* .

Herel, D. & Mikolov, T. (2023), 'Advancing state of the art in language modeling', *arXiv preprint arXiv:2312.03735* .

Karthik, S., Coskun, H., Akata, Z., Tulyakov, S., Ren, J. & Kag, A. (2024), 'Scalable ranked preference optimization for text-to-image generation', *arXiv preprint arXiv:2410.18013* .

Ke, Z., Xu, A., Ming, Y., Nguyen, X.-P., Xiong, C. & Joty, S. (2025), 'Mas-zero: Designing multi-agent systems with zero supervision', *arXiv preprint arXiv:2505.14996* .

Kim, T., Singh, J., Mehri, S., Acikgoz, E. C., Mukherjee, S., Bozdag, N. B., Shashidhar, S. & Tur, G. (2025), 'Pipa: A unified evaluation protocol for diagnosing interactive planning agents', *arXiv preprint arXiv:2505.01592* .

Krishnan, N. (2025*a*), 'Advancing multi-agent systems through model context protocol: Architecture, implementation, and applications', *arXiv preprint arXiv:2504.21030* .

Krishnan, N. (2025*b*), 'Ai agents: Evolution, architecture, and real-world applications', *arXiv preprint arXiv:2503.12687* .

Lee, J. & Hockenmaier, J. (2025), 'Evaluating step-by-step reasoning traces: A survey', *arXiv preprint arXiv:2502.12289* .

Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A. & Jiang, Y. (2024), 'From generation to judgment: Opportunities and challenges of llm-as-a-judge', *arXiv preprint arXiv:2411.16594* .

Li, J., Sun, S., Yuan, W., Fan, R.-Z., Zhao, H. & Liu, P. (2023), 'Generative judge for evaluating alignment', *arXiv preprint arXiv:2310.05470* .

Li, Z., Wang, C., Ma, P., Wu, D., Wang, S., Gao, C. & Liu, Y. (2023), 'Split and merge: Aligning position biases in large language model based evaluators', *arXiv preprint arXiv:2310.01432* .

Li, Z., Xu, X., Shen, T., Xu, C., Gu, J.-C. & Tao, C. (2024), 'Leveraging large language models for nlg evaluation: A survey', *arXiv preprint arXiv:2401.07103* .

Liu, B., Li, X., Zhang, J., Wang, J., He, T., Hong, S. et al. (2025), 'Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems', *arXiv preprint arXiv:2504.01990* .

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R. & Zhu, C. (2023), 'G-eval: Nlg evaluation using gpt-4 with better human alignment', *arXiv preprint arXiv:2303.16634* .

Ma, C., Zhang, J., Zhu, Z., Yang, C., Yang, Y., Jin, Y., Lan, Z. & Kong, L. (2024), 'Agentboard: An analytical evaluation board of multi-turn llm agents', *arXiv preprint arXiv:2401.13178* .

Maharana, A., Kamath, A., Clark, C., Bansal, M. & Kembhavi, A. (2023), 'Exposing and addressing cross-task inconsistency in unified vision-language models', *arXiv preprint arXiv:2303.16133* .

Marjanović, S. V., Patel, A., Adlakha, V., Aghajohari, M., BehnamGhader, P., Bhatia, M., Khandelwal, A. & Kraft, A. (2025), 'Deepseek-r1 thoughtology: Let's <think> about llm reasoning', *arXiv preprint arXiv:2504.07128* .

Masterman, T., Besen, S., Sawtell, M. & Chao, A. (2024), 'The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey', *arXiv preprint arXiv:2404.11584* .

Muldrew, W., Hayes, P., Zhang, M. & Barber, D. (2024), 'Active preference learning for large language models', *arXiv preprint arXiv:2402.08114* .

Murakami, S., Hoshino, S. & Zhang, P. (2023), 'Natural language generation for advertising: A survey', *arXiv preprint arXiv:2306.12719* .

Ni, J., Song, Y., Ghosal, D., Li, B., Zhang, D. J., Yue, X., Xue, F. & Zheng, Z. (2024), 'Mixeval-x: Any-to-any evaluations from real-world data mixtures', *arXiv preprint arXiv:2410.13754* .

Patil, A. (2025), 'Advancing reasoning in large language models: Promising methods and approaches', *arXiv preprint arXiv:2502.03671* .

Pauli, A. B., Augenstein, I. & Assent, I. (2024), 'Measuring and benchmarking large language models' capabilities to generate persuasive language', *arXiv preprint arXiv:2406.17753* .

Peng, S., Wang, W., Tian, Z., Yang, S., Wu, X., Xu, H., Zhang, C. & Isobe, T. (2025), 'Omni-dpo: A dual-perspective paradigm for dynamic preference learning of llms', *arXiv preprint arXiv:2506.10054* .

Pimentel, M. A., Christophe, C., Raha, T., Munjal, P., Kanithi, P. K. & Khan, S. (2024), 'Beyond metrics: A critical analysis of the variability in large language model evaluation frameworks', *arXiv preprint arXiv:2407.21072* .

Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C. & Huang, F. (2022), 'Reasoning with language model prompting: A survey', *arXiv preprint arXiv:2212.09597* .

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D. & Finn, C. (2023), 'Direct preference optimization: Your language model is secretly a reward model', *arXiv preprint arXiv:2305.18290* .

Rony, M. R. A. H., Kovriguina, L., Chaudhuri, D., Usbeck, R. & Lehmann, J. (2022), 'Rome: A robust metric for evaluating natural language generation', *arXiv preprint arXiv:2203.09183* .

Sapkota, R., Roumeliotis, K. I. & Karkee, M. (2025), 'Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenge', *arXiv preprint arXiv:2505.10468* .

Schmidtová, P., Mahamood, S., Balloccu, S., Dušek, O., Gatt, A., Gkatzia, D., Howcroft, D. M. & Plátek, O. (2024), 'Automatic metrics in natural language generation: A survey of current evaluation practices', *arXiv preprint arXiv:2408.09169* .

Seth, P. & Sankarapu, V. K. (2025), 'Bridging the gap in xai-why reliable metrics matter for explainability and compliance', *arXiv preprint arXiv:2502.04695* .

Shao, Z., Gong, Y., Shen, Y., Huang, M., Duan, N. & Chen, W. (2023), 'Synthetic prompting: Generating chain-of-thought demonstrations for large language models', *arXiv preprint arXiv:2302.00618* .

Siegel, Z. S., Kapoor, S., Nagdir, N., Stroebl, B. & Narayanan, A. (2024), 'Core-bench: Fostering the credibility of published research through a computational reproducibility agent benchmark', *arXiv preprint arXiv:2409.11363* .

Sui, Y., Chuang, Y.-N., Wang, G., Zhang, J., Zhang, T., Yuan, J., Liu, H. & Wen, A. (2025), 'Stop overthinking: A survey on efficient reasoning for large language models', *arXiv preprint arXiv:2503.16419* .

Talebirad, Y. & Nadiri, A. (2023), 'Multi-agent collaboration: Harnessing the power of intelligent llm agents', *arXiv preprint arXiv:2306.03314* .

Urlana, A., Kumar, C. V., Singh, A. K., Garlapati, B. M., Chalamala, S. R. & Mishra, R. (2024), 'Llms with industrial lens: Deciphering the challenges and prospects – a survey', *arXiv preprint arXiv:2402.14558* .

Wang, Q., Lou, Z., Tang, Z., Chen, N., Zhao, X., Zhang, W., Song, D. & He, B. (2025), 'Assessing judging bias in large reasoning models: An empirical study', *arXiv preprint arXiv:2504.09946* .

Wang, R., Sun, J., Hua, S. & Fang, Q. (2024), 'Asft: Aligned supervised fine-tuning through absolute likelihood', *arXiv preprint arXiv:2409.10571* .

Xu, A., Bansal, S., Ming, Y., Yavuz, S. & Joty, S. (2025), 'Does context matter? contextualjudgebench for evaluating llm-based judges in contextual settings', *arXiv preprint arXiv:2503.15620* .

Xu, X., Tao, C., Shen, T., Xu, C., Xu, H., Long, G. & guang Lou, J. (2023), 'Re-reading improves reasoning in language models', *arXiv preprint arXiv:2309.06275* .

Yan, B., Zhang, X., Zhang, L., Zhang, L., Zhou, Z., Miao, D. & Li, C. (2025), 'Beyond self-talk: A communication-centric survey of llm-based multi-agent systems', *arXiv preprint arXiv:2502.14321* .

Yang, H., Bao, R., Xiao, C., Ma, J., Bhatia, P., Gao, S. & Kass-Hout, T. (2025), 'Any large language model can be a reliable judge: Debiasing with a reasoning-based bias detector', *arXiv preprint arXiv:2505.17100* .

Ye, J., Wang, Y., Huang, Y., Chen, D., Zhang, Q., Moniz, N., Gao, T. & Geyer, W. (2024), 'Justice or prejudice? quantifying biases in llm-as-a-judge', *arXiv preprint arXiv:2410.02736* .

Yehudai, A., Eden, L., Li, A., Uziel, G., Zhao, Y., Bar-Haim, R., Cohan, A. & Shmueli-Scheuer, M. (2025), 'Survey on evaluation of llm-based agents', *arXiv preprint arXiv:2503.16416* .

Zeng, Z., Chen, P., Jiang, H. & Jia, J. (2023), 'Challenge llms to reason about reasoning: A benchmark to unveil cognitive depth in llms', *arXiv preprint arXiv:2312.17080* .

Zhang, R. & Tetreault, J. (2019), 'This email could save your life: Introducing the task of email subject line generation', *arXiv preprint arXiv:1906.03497* .

Zhang, Z., Zheng, C., Tang, D., Sun, K., Ma, Y., Bu, Y., Zhou, X. & Zhao, L. (2023), 'Balancing specialized and general skills in llms: The impact of modern tuning and data strategy', *arXiv preprint arXiv:2310.04945* .

Zheng, C., Ke, P., Zhang, Z. & Huang, M. (2023), 'Click: Controllable text generation with sequence likelihood contrastive learning', *arXiv preprint arXiv:2306.03350* .

Zhou, P., Feng, Y., Julaiti, H. & Yang, Z. (2025), 'Why do ai agents communicate in human language?', *arXiv preprint arXiv:2506.02739* .

Zhu, K., Du, H., Hong, Z., Yang, X., Guo, S., Wang, Z., Wang, Z. & Qian, C. (2025), 'Multiagentbench: Evaluating the collaboration and competition of llm agents', *arXiv preprint arXiv:2503.01935* .

Zhu, K., Wang, J., Zhao, Q., Xu, R. & Xie, X. (2024), 'Dynamic evaluation of large language models by meta probing agents', *arXiv preprint arXiv:2402.14865* .

# Appendices

# Appendix A

# Experimental Setup Details

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu.

Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.