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 7, 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

One or two sentences providing a basic introduction to the field, comprehensible to a scientist in any discipline. Two to three sentences of more detailed background, comprehensible to scientists in related disciplines. One sentence clearly stating the general problem being addressed by this particular study. One sentence summarising the main result (with the words "here I show" or their equivalent). Two or three sentences explaining what the main result reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge. One or two sentences to put the results into a more general context. Two or three sentences to provide a broader perspective, readily comprehensible to a scientist in any discipline.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

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.

Chapter 2

Literature Review

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.

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).

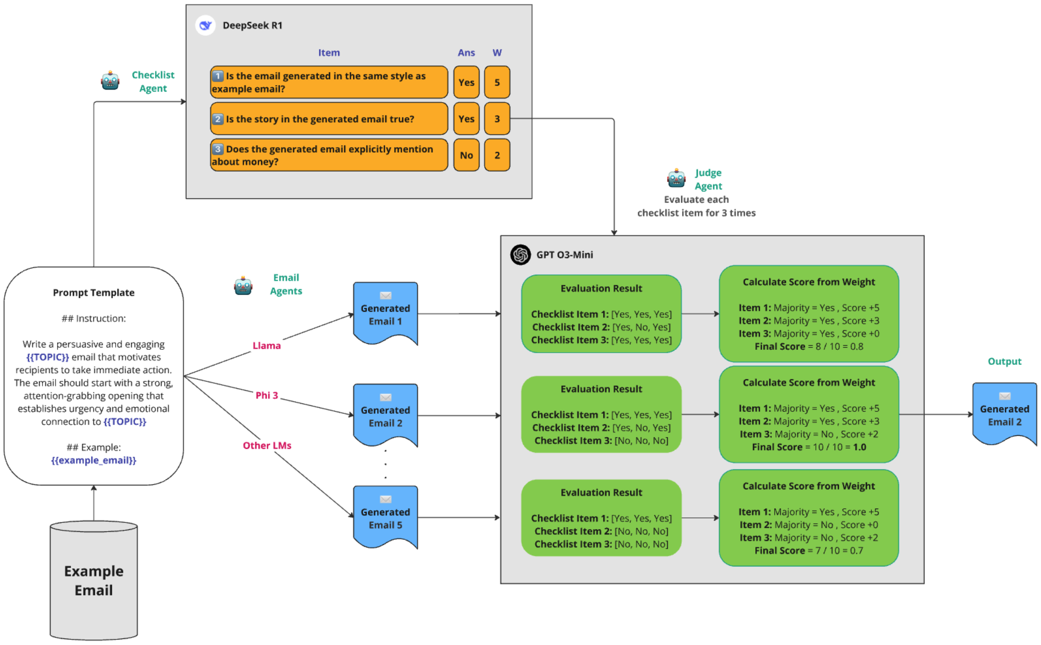
3.1.2 Three-Agent Architecture Design

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.



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}}

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

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

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.

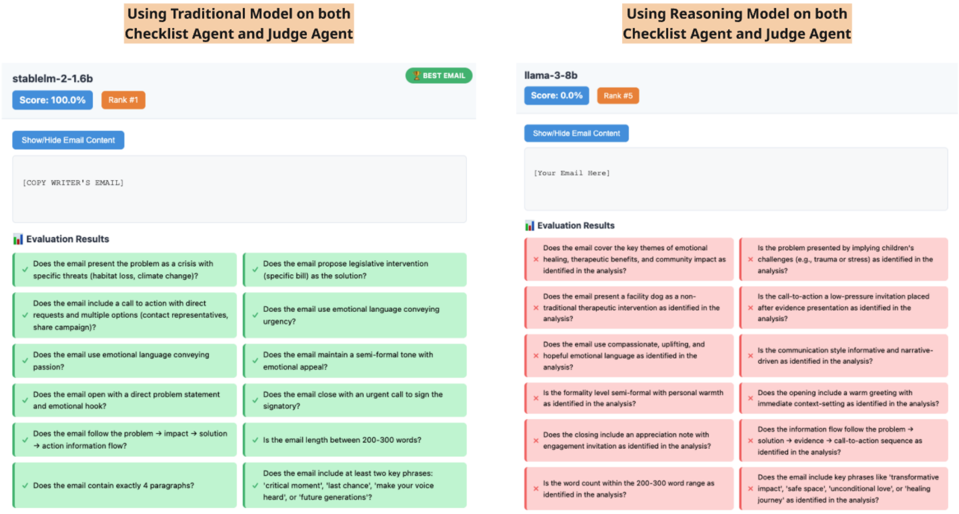
3.2.5 Agent Model Experimentation and Selection

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.

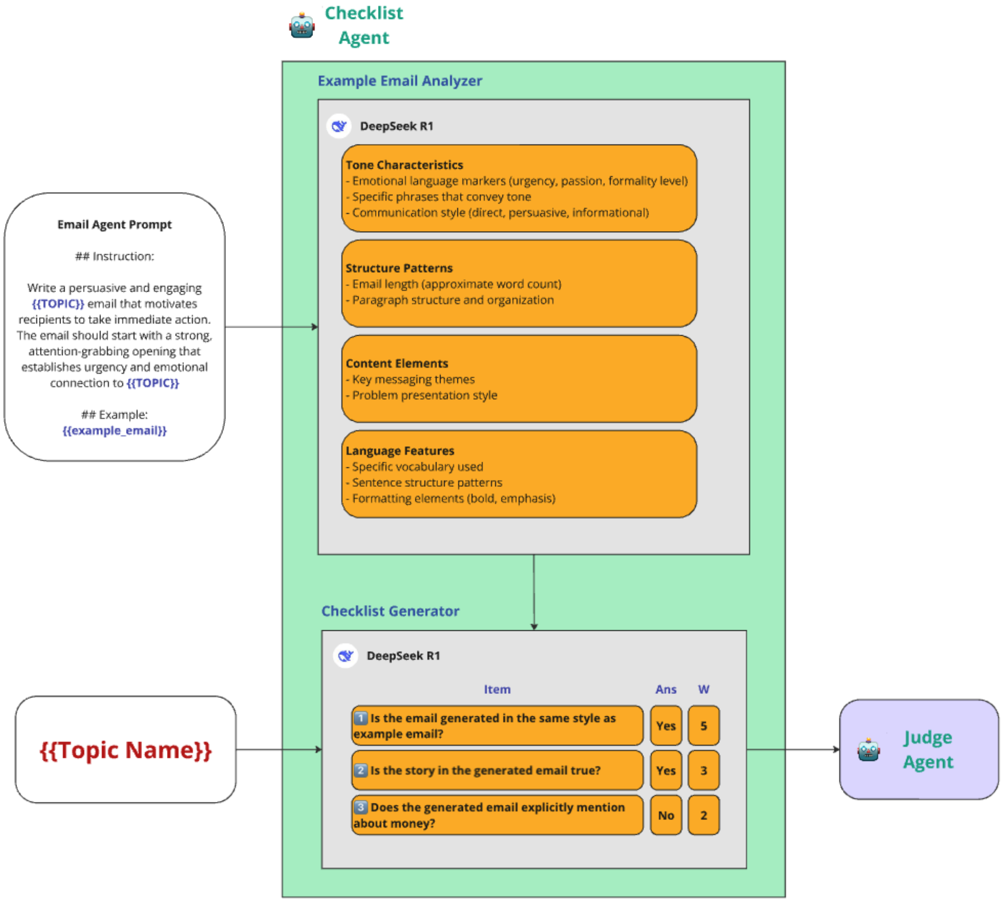
Computational efficiency analysis revealed that the Hybrid prompting strategy achieved significant reduction in processing overhead compared to Full-Prompt analysis while maintaining high evaluation quality. This efficiency gain, combined with improved reasoning model focus, enables practical implementation 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
## 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}}
Checklist Generator
DeepSeek R1
{{Topic Name}}
Judge Agent

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 T0026-T0100	1 Human + 5 AI 5 AI-generated	5 4	125 300
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.

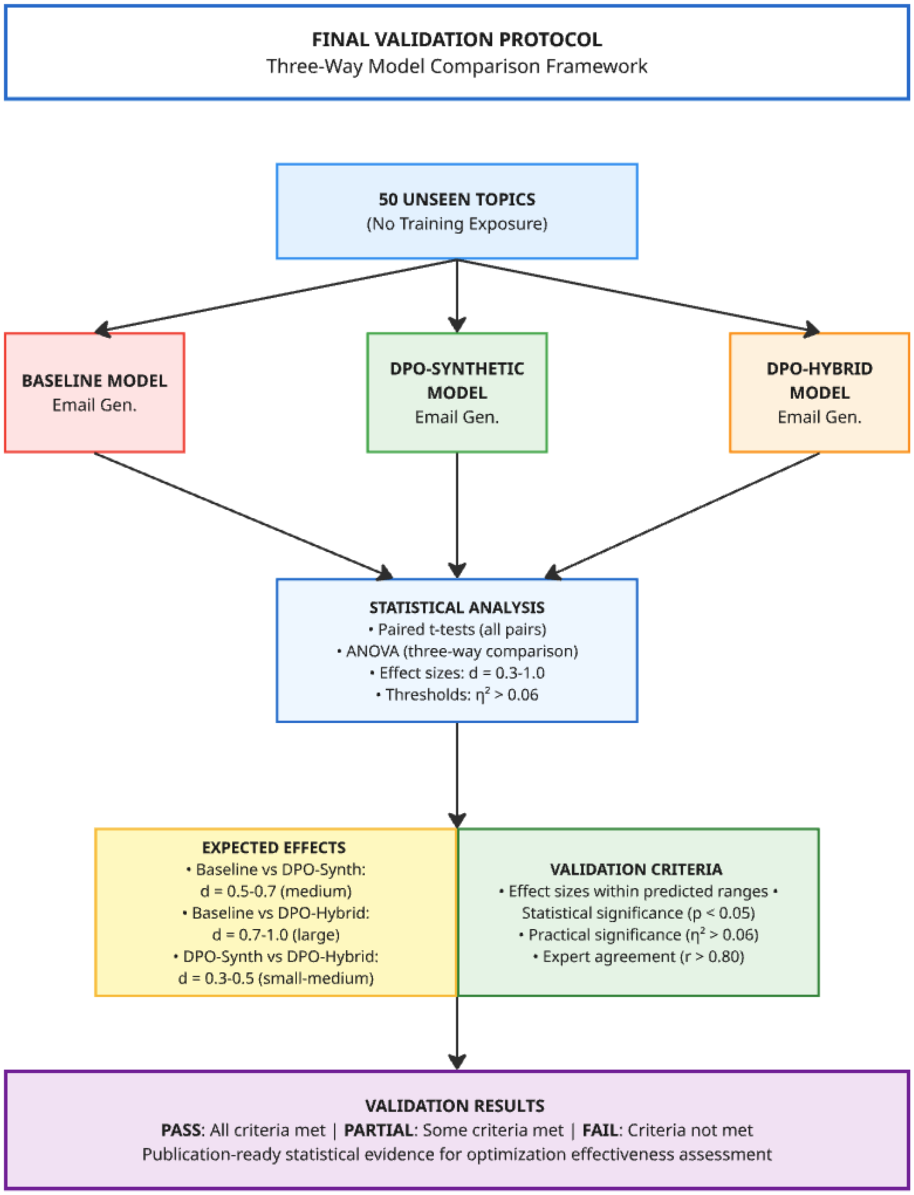

FINAL VALIDATION PROTOCOL
Three-Way Model Comparison Framework
50 UNSEEN TOPICS
(No Training Exposure)
BASELINE MODEL
Email Gen.
DPO-SYNTHETIC MODEL
Email Gen.
DPO-HYBRID MODEL
Email Gen.
STATISTICAL ANALYSIS
• Paired t-tests (all pairs)
• ANOVA (three-way comparison)
• Effect sizes: d = 0.3-1.0
• Thresholds: 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).

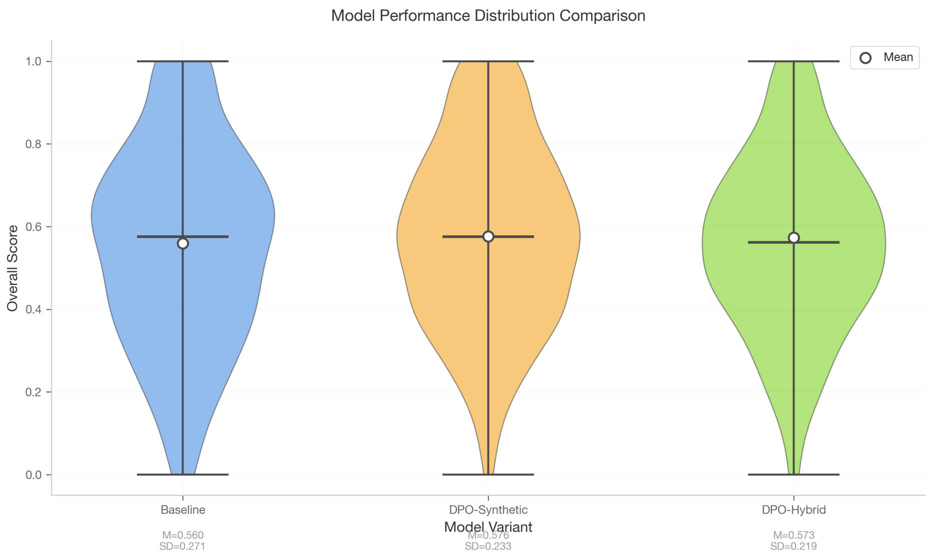


Figure 4.1: *Model Performance Comparison. Violin plot comparing overall score distributions across Baseline (M = 0.560, SD = 0.271), DPO-Synthetic (M = 0.576, SD = 0.233), and DPO-Hybrid (M = 0.573, SD = 0.219) variants. Violin plots show distribution shapes with kernel density estimation, medians, 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, η^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.

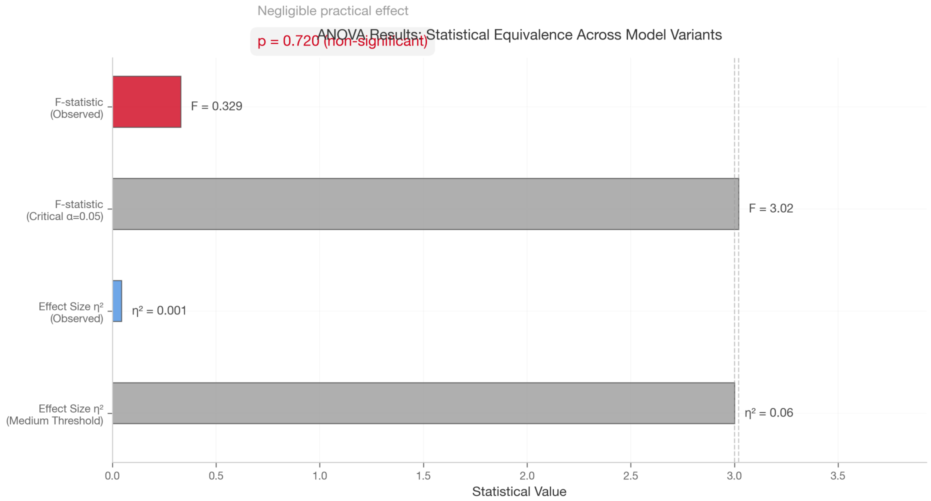


Figure 4.2: *ANOVA Results Summary. Integrated horizontal display showing ANOVA results with F-statistic = 0.329 (well below critical threshold of 3.02) and* $\eta^2 = 0.001$ *(well below medium effect threshold of 0.06). Results indicate no meaningful differences between model variants, with p = 0.720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

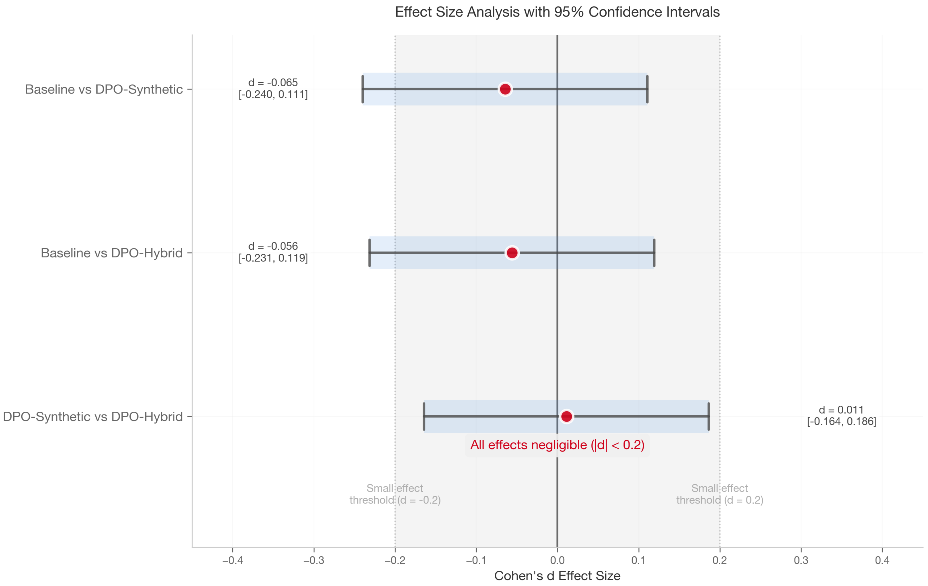


Figure 4.3: *Effect Size Forest Plot with 95% Confidence Intervals. Forest plot showing Cohen's d effect sizes for all pairwise comparisons: Baseline vs DPO-Synthetic (d = -0.065 [-0.240, 0.111]), Baseline vs DPO-Hybrid (d = -0.056 [-0.231, 0.119]), and DPO-Synthetic vs DPO-Hybrid (d = 0.011 [-0.164, 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).

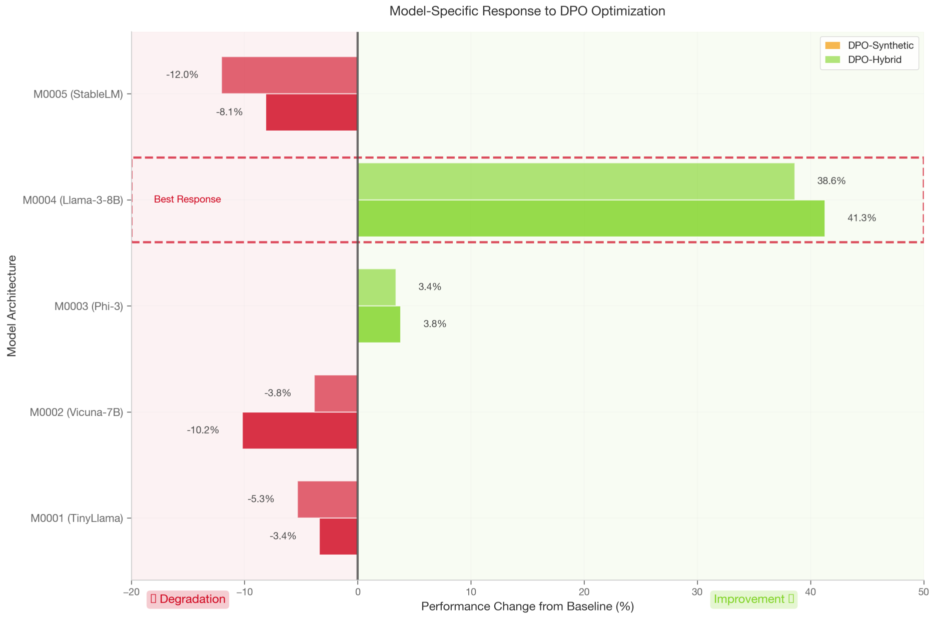


Figure 4.4: *Model-Specific Improvement Forest Plot. Forest plot showing improvement rates for each individual model across DPO-Synthetic and DPO-Hybrid variants compared to baseline. Model M0004 (Llama-3-8B) demonstrates the largest improvements (+41.3% Synthetic, +38.6% Hybrid), while M0001 (TinyLlama), 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.*

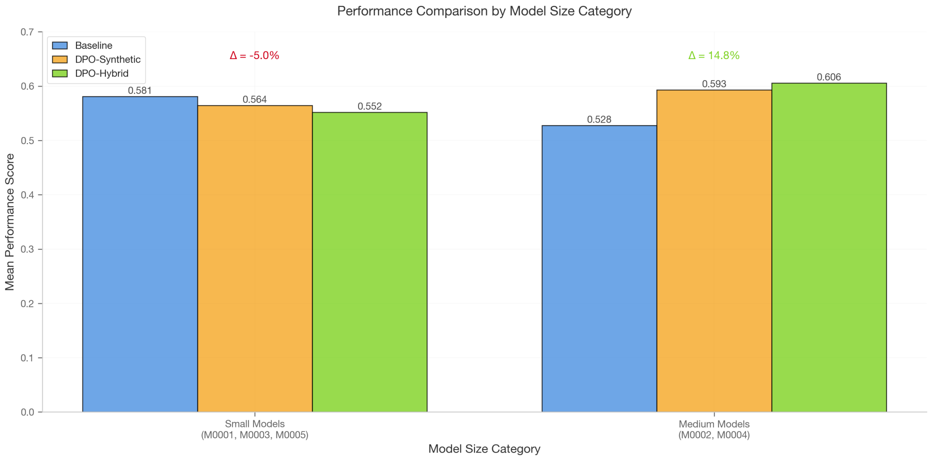


Figure 4.5: *Model Size Group Performance Comparison. Performance comparison by model size groups showing aggregated performance within small models (M0001, M0003, M0005) and medium models (M0002, M0004). Small models show baseline mean = 0.581, DPO-Synthetic mean = 0.564, DPO-Hybrid mean = 0.552. Medium 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 η^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 4.6). 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.

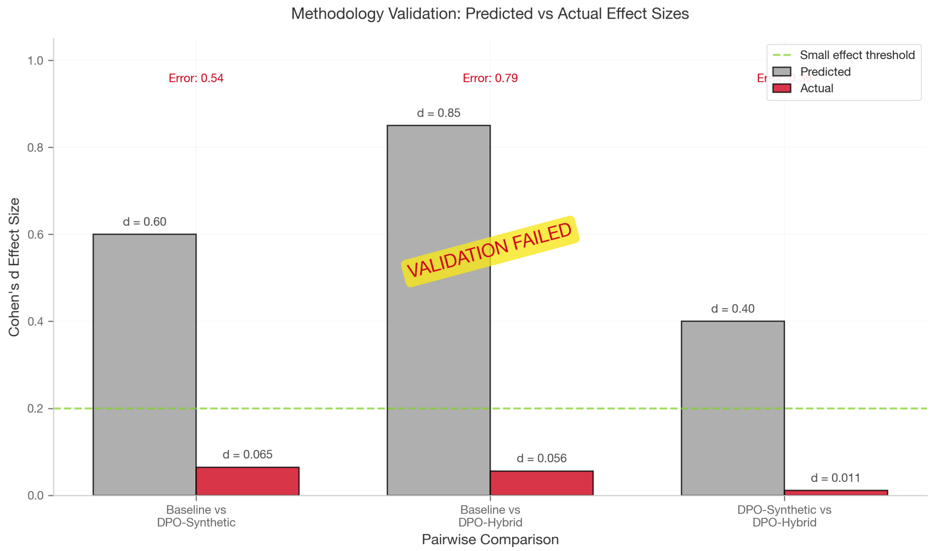


Figure 4.6: *Methodology Validation: Predicted vs Actual Effect Sizes. Comparison of predicted versus actual effect sizes showing systematic validation failure across all theoretical predictions. The largest discrepancy occurred in the Baseline vs DPO-Hybrid comparison (predicted d = 0.85, actual d = -0.056), representing a prediction error exceeding 0.9 effect size units. All predictions substantially overestimated actual empirical effects, indicating fundamental limitations in the theoretical framework.*

Chapter 5

Discussion

5.1 Research Impact and Contributions

The methodology establishes foundations for several important research and practical contributions to the field of automated content generation. Research impact includes the development of novel multi-agent evaluation frameworks applicable to other content generation domains, validation of consistency sampling methodologies for reliable model assessment, and demonstration of DPO fine-tuning effectiveness in domain-specific content generation tasks.

5.2 Future Research Directions

Future research directions emerging from this methodology include extension to other persuasive communication domains such as marketing and advocacy campaigns, investigation of cross-cultural effectiveness in fundraising email generation across different geographic and cultural contexts, and development of real-time adaptation mechanisms that adjust generation strategies based on audience response feedback.

Table 5.1: *Future research directions and methodological extensions*

Research Area	Proposed Extension	Expected Impact
[Future research directions table to be completed]		

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.

Bibliography

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* .

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* .

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

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* .

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* .

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

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.