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

July 20, 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 is structured in three stages: Stage 1 (System Design and Setup) establishes the foundational framework, Stage 2 (Experimental Implementation) details the execution procedures, and Stage 3 (Enhancement and Analysis) describes the analytical approach and planned extensions including Direct Preference Optimization fine-tuning.

3.1 Research Design and Approach

This study adopts a quantitative comparative research paradigm to systematically evaluate the performance of different language models in automated email generation tasks. The research is grounded in experimental design principles with controlled variables and systematic evaluation procedures to ensure methodological rigor and reproducible results.

The central research problem addresses the effectiveness of various language model architectures and sizes in generating high-quality fundraising emails within a structured evaluation framework. This investigation is motivated by the growing need for automated content generation systems that can produce contextually appropriate and persuasive communication while maintaining consistency and quality across different model implementations Murakami et al. (2023), Zheng et al. (2023).

The methodological approach employs a multi-agent system design as a novel contribution to the field of automated text generation evaluation Guo et al. (2024), Yan et al. (2025). Unlike traditional single-model assessment approaches, this methodology introduces specialist agents for distinct phases of the evaluation process, enabling more comprehensive and systematic comparison of model capabilities Yehudai et al. (2025). The multi-agent approach provides several advantages over conventional evaluation methods: enhanced objectivity through agent specialization, systematic evaluation criteria generation, and standardized assessment protocols across all tested models Ma et al. (2024).

The research questions guiding this investigation focus on comparative performance assessment across different model categories, consistency of output quality within individual models, and the effectiveness of the proposed multi-agent evaluation framework in providing reliable and valid assessments of generated content quality.

3.2 System Architecture Overview

The proposed system implements a three-agent architecture designed to systematically evaluate language model performance in email generation tasks. Each agent serves a distinct function within the evaluation pipeline, ensuring comprehensive assessment while maintaining methodological consistency across all experimental conditions.

The **Email Generator Agent** serves as the primary content creation component, responsible for generating fundraising emails based on standardized prompts and topic specifications.

This agent interfaces with multiple language models sequentially, ensuring consistent input conditions while capturing the unique characteristics and capabilities of each model under evaluation.

The **Checklist Creator Agent** functions as the evaluation criteria development component, generating structured assessment frameworks for each generated email. This agent utilizes DeepSeek R1, a reasoning-capable language model specifically selected for its analytical capabilities in evaluation criteria development. The agent produces binary evaluation checklists with priority weighting, ensuring that assessment criteria are both comprehensive and relevant to the specific content and context of each generated email. The checklist generation process maintains consistency in evaluation standards while adapting to the nuanced characteristics of different email content.

The **Judge Agent** operates as the performance assessment and ranking component, applying the generated checklists to evaluate email quality systematically. This agent employs GPT O3 Mini, selected for its advanced reasoning capabilities and consistency in evaluation tasks. The agent implements a probability-based scoring methodology that accounts for both binary assessment outcomes and priority weighting, providing quantitative measures for comparative analysis across different models and topics.

Figure 3.1: *Three-agent system architecture with reasoning-capable models showing agent interactions and data flow*

The multi-model orchestration strategy enables parallel processing of different language models while maintaining experimental control and consistency. This approach maximizes computational efficiency while ensuring that each model receives identical input conditions and evaluation procedures, thereby supporting valid comparative analysis across the full range of tested models.

3.3 Model Selection and Categorization

The model selection process follows a systematic taxonomy based on parameter count and architectural characteristics, ensuring representative coverage across the spectrum of available open-source language models. This categorization enables meaningful comparison both within and across model size categories while accounting for the diverse capabilities and computational requirements of different model architectures.

3.3.1 Model Taxonomy and Categories

Models are categorized into three primary groups based on parameter count and intended use cases:

Small Models (1.1B-1.6B parameters) focus on resource efficiency and rapid inference capabilities. These models represent the lower bound of contemporary language model capabilities while offering practical advantages in computational requirements and deployment feasibility. The inclusion of small models enables assessment of whether compact architectures can achieve acceptable performance in structured email generation tasks.

Medium Models (7B-8B parameters) represent a balance between performance capabilities and computational efficiency. This category encompasses models that demonstrate

substantial language understanding and generation capabilities while remaining accessible for practical deployment scenarios. Medium models serve as the primary comparison baseline, representing the current mainstream approach to language model deployment.

Large Models (34B-70B parameters) provide assessment of maximum capability within the current open-source model landscape. These models enable evaluation of whether increased parameter count translates to proportional improvements in email generation quality and consistency, while establishing upper bounds for performance expectations within the experimental framework.

Reasoning Models represent a specialized category of language models optimized for analytical and evaluation tasks. This category includes models specifically designed for logical reasoning, consistency in evaluation, and systematic analysis capabilities. The integration of reasoning models addresses the specific requirements of evaluation agents within the multi-agent framework.

The research employs a systematic Unique Identifier (UID) system for model tracking and experimental organization. Models M0001 through M0007 represent the primary email generation models categorized by size, while reasoning models (DeepSeek R1, GPT O3 Mini) serve specialized evaluation functions. Additionally, Direct Preference Optimization (DPO) variants are available for models M0001-M0005, enabling comparative analysis between baseline and preference-optimized versions of the same architectural foundations.

Table 3.1: *Enhanced Language Model Specifications with Dual-Method DPO Variants*

UID	Model Name	Parameters	Category	DPO-Synthetic	DPO-Hybrid	Primary Use
M0001	TinyLlama-1.1B	1.1B	Small	Available	Available	Email Generat
M0002	Vicuna-7B	7B	Medium	Available	Available	Email Generat
M0003	Phi-3-Mini	3.8B	Small	Available	Available	Email Generat
M0004	Llama-3-8B	8B	Medium	Available	Available	Email Generat
M0005	StableLM-2-1.6B	1.6B	Small	Available	Available	Email Generat
M0006	Yi-34B	34B	Large	N/A	N/A	Evaluation Ta
M0007	Llama-3-70B	70B	Large	N/A	N/A	Evaluation Ta
-	DeepSeek R1	8B	Reasoning	N/A	N/A	Checklist Gen
-	GPT O3 Mini	-	Reasoning	N/A	N/A	Evaluation/Ju

3.3.2 Selection Criteria and Rationale

The model selection process prioritizes open-source implementations to ensure reproducibility and accessibility of the research findings. Selection criteria include architectural diversity to capture different approaches to language modeling, availability of appropriate quantization options for efficient deployment, and demonstrated performance in text generation tasks based on existing literature and benchmarks.

Diversity considerations encompass different transformer architectures, training methodologies, and fine-tuning approaches represented across the selected models. This diversity ensures that the evaluation captures fundamental differences in model design and training rather than minor variations within a single architectural family.

3.3.3 Agent Model Optimization

The selection of specific models for the Checklist Creator and Judge agents follows a systematic optimization process based on preliminary empirical testing of reasoning capabilities across different model architectures. This optimization process represents a methodological advancement that prioritizes reasoning-capable models for evaluation tasks requiring analytical depth and consistency.

Performance Criteria and Selection Rationale

The selection criteria for agent-specific models prioritize reasoning capabilities, evaluation consistency, and analytical depth over general text generation performance. Reasoning-capable models demonstrate superior performance in structured evaluation tasks through enhanced logical analysis capabilities, improved consistency across repeated evaluations, systematic approach to criteria development, and reduced bias in comparative assessments.

The empirical evidence supporting reasoning model superiority in evaluation tasks (detailed in Section 4.1) provides methodological justification for the agent-specific model selection approach. This optimization strategy ensures that each agent operates with the most appropriate model architecture for its designated function within the multi-agent evaluation framework.

Figure 3.2: *Agent Model Selection Comparison: Reasoning vs Traditional Models*

3.4 Dataset and Topic Selection

The selection of charity fundraising emails as the evaluation domain provides several methodological advantages: clear assessment criteria for persuasive and contextually appropriate content, well-defined audience expectations and communication goals, and sufficient complexity to differentiate between model capabilities while remaining accessible for systematic evaluation Zhang and Tetreault (2019), Pauli et al. (2024).

3.4.1 Topic Development and Validation

The experimental dataset comprises 100 distinct fundraising topics distributed across 12 charity categories, providing comprehensive coverage of the fundraising domain while ensuring sufficient sample size for statistical analysis. Topic development follows a systematic process beginning with charity sector analysis and stakeholder consultation to identify representative fundraising scenarios. The expanded dataset enables dual-method DPO experimentation through strategic integration of human-authored content with systematic synthetic data generation.

The 12 charity categories include healthcare and medical research, education and youth development, environmental conservation, humanitarian aid and disaster relief, animal welfare, poverty alleviation and social services, elderly care and support, community development, disability support and accessibility, mental health awareness, refugee assistance, and

emergency medical services. This categorization ensures coverage of major charitable sectors while providing sufficient topic diversity for robust model evaluation.

Table 3.2: *Distribution of fundraising topics across charity categories*

Category	Topic Count	Examples
[Topic distribution table to be completed]		

3.4.2 Content Validation and Standardization

Topic standardization procedures ensure consistency in complexity, scope, and evaluation criteria across all experimental conditions. Each topic undergoes validation through expert review processes involving fundraising professionals and communication specialists to verify authenticity and appropriateness of the fundraising scenarios.

Content validation addresses both factual accuracy and representativeness of real-world fundraising communications. The validation process includes review of topic descriptions for clarity and specificity, assessment of fundraising goal appropriateness and realism, evaluation of target audience definition and communication objectives, and verification of ethical considerations and sensitivity requirements.

Ethical considerations in domain selection include ensuring respectful representation of charitable causes, avoiding exploitation of sensitive social issues for research purposes, and maintaining awareness of the potential impact of generated content on public perception of charitable organizations and causes.

The standardized topic framework provides consistent input conditions for all models while allowing sufficient variation to assess adaptability and contextual understanding across different fundraising scenarios. This approach supports both within-model consistency analysis and cross-model comparative evaluation within a controlled experimental environment.

3.4.3 Human Baseline Integration

The methodology incorporates a systematic human baseline integration approach that strategically combines authentic human-authored fundraising emails with the synthetic experimental framework. This integration serves dual purposes: establishing performance benchmarks through comparison with human-generated content and providing high-quality preference data for hybrid DPO training procedures.

Human Email Collection and Validation

The human baseline dataset comprises 25 high-quality fundraising emails corresponding to topics T0001-T0025, collected through systematic collaboration with fundraising professionals and charitable organizations. Email selection follows rigorous criteria designed to ensure representativeness and quality: demonstrated effectiveness in actual fundraising campaigns, adherence to professional communication standards, coverage of diverse charitable sectors within the first 25 topic categories, and compliance with ethical guidelines for research use.

Human email validation employs a comprehensive multi-stage process that includes expert review by fundraising professionals to assess quality and effectiveness, content analysis to

verify alignment with corresponding synthetic topic specifications, ethical review to ensure appropriate representation of charitable causes, and linguistic analysis to confirm professional communication standards. This validation process ensures that human examples represent authentic best practices in fundraising communication while maintaining compatibility with the experimental framework.

Strategic Topic Selection for Human Data

The selection of the first 25 topics (T0001-T0025) for human email integration follows systematic criteria designed to maximize experimental validity while maintaining practical feasibility. The selection ensures balanced representation across the 12 charity categories, sufficient diversity to assess generalization capabilities, practical feasibility for human email collection efforts, and strategic positioning to enable meaningful hybrid preference learning.

This strategic approach enables the research to assess fundamental questions regarding the value of human expertise in preference learning while maintaining sufficient synthetic data coverage for comprehensive model evaluation. The 25-75 ratio of human-to-synthetic preference data provides an optimal balance between authentic human insight and scalable synthetic data generation **?**.

Human-Synthetic Data Integration Methodology

The integration methodology implements systematic procedures to ensure compatibility between human-authored and synthetic preference data while preserving the distinct characteristics of each data source. Integration procedures include normalization of formatting standards to ensure consistency across human and synthetic examples, quality threshold alignment to maintain comparable standards between data sources, topic-specific calibration to verify contextual consistency within individual categories, and statistical validation to confirm that human examples represent meaningful quality improvements.

The methodology addresses potential challenges in human-synthetic integration through careful attention to evaluation consistency, comparative quality assessment, and systematic documentation of integration procedures. This approach ensures that hybrid preference learning achieves optimal effectiveness while maintaining experimental validity and reproducibility standards.

Table 3.3: *Human-Synthetic Data Integration Strategy for Dual-Method DPO*

Topic Range	**Data Source**	**Count**	**Purpose**
T0001-T0025	Human-authored	25	Hybrid preference learning
T0026-T0100	Synthetic generation	75	Scalable preference data
All Topics	Judge Agent scores	100	Synthetic preference learning
Total Preference Pairs	**DPO-Synthetic: 100**	**DPO-Hybrid: 100**	**Comparison Framework**

3.5 Multi-Modal Checklist Generation Framework

This research introduces a novel multi-modal checklist generation framework that systematically evaluates different approaches to evaluation criteria development. The framework

implements three distinct operational modes, each designed to test specific aspects of checklist generation effectiveness and computational efficiency. This methodological innovation enables comprehensive analysis of the trade-offs between analytical depth, processing efficiency, and evaluation quality.

3.5.1 Three-Mode Experimental Design

The multi-modal framework employs three systematically designed operational modes that represent different approaches to evaluation criteria generation. Each mode implements distinct processing strategies while maintaining consistency in output format and evaluation standards.

Full-Prompt Mode

The Full-Prompt mode represents the comprehensive analytical approach, providing complete contextual analysis with access to the entire email content, topic specifications, and detailed prompt instructions. This mode enables the Checklist Creator Agent to perform holistic assessment of email content, considering all available contextual information for evaluation criteria development.

The Full-Prompt mode serves as the methodological baseline, representing the optimal conditions for checklist generation when computational resources and processing time are not constraining factors. This mode enables assessment of maximum possible evaluation quality achievable through comprehensive contextual analysis.

Extract-Only Mode

The Extract-Only mode implements minimal context processing for efficiency comparison, utilizing only essential content elements and abbreviated prompt instructions. This mode tests the hypothesis that effective evaluation criteria can be generated with reduced computational overhead while maintaining acceptable evaluation quality standards.

The design of Extract-Only mode prioritizes processing efficiency while preserving core evaluation functionality. This approach enables assessment of the minimum viable approach to checklist generation, providing insights into the essential components required for effective evaluation criteria development.

Hybrid Mode

The Hybrid mode implements a two-step systematic analysis that combines extraction and processing phases in a structured analytical framework. This mode represents an intermediate approach that balances comprehensive analysis with processing efficiency through systematic decomposition of the evaluation task.

The Hybrid mode methodology involves initial content extraction followed by structured analysis of extracted elements. This approach tests whether systematic two-phase processing can achieve evaluation quality comparable to comprehensive analysis while maintaining improved processing efficiency relative to the Full-Prompt mode.

3.5.2 Methodological Justification and Comparative Framework

The three-mode design enables systematic evaluation of fundamental questions regarding evaluation criteria generation: the relationship between analytical depth and evaluation quality, the impact of processing efficiency constraints on assessment accuracy, and the effectiveness of structured analytical approaches in balancing quality and efficiency considerations.

Each mode tests distinct hypotheses about optimal approaches to evaluation criteria development. The Full-Prompt mode tests the upper bounds of evaluation quality achievable through comprehensive analysis. The Extract-Only mode tests the minimal viable approach to criteria generation. The Hybrid mode tests whether systematic structured analysis can optimize the quality-efficiency trade-off.

The comparative framework enables systematic assessment of mode performance across multiple evaluation dimensions, including criteria quality, consistency reliability, processing efficiency, and evaluation accuracy. This multi-dimensional analysis approach provides comprehensive insights into the strengths and limitations of each operational mode.

Figure 3.3: *Three-Mode Checklist Generation Framework showing workflow differences and processing approaches*

Table 3.4: *Mode-Specific Parameters and Expected Outcomes*

Mode	Context Level	Processing Steps	Expected Quality	Efficiency
Full-Prompt	Complete	Single-phase	High	Low
Extract-Only	Minimal	Single-phase	Medium	High
Hybrid	Structured	Two-phase	High-Medium	Medium

3.6 Experimental Design

The experimental design implements a comprehensive multi-phase experimental protocol that systematically evaluates language model performance across diverse fundraising contexts while enabling direct comparison of baseline and dual-method preference-optimized model variants. This advanced design enables systematic comparison of model capabilities both within individual model categories and across the full spectrum of tested architectures, while providing quantitative assessment of dual-method Direct Preference Optimization effectiveness.

3.6.1 Multi-Phase Experimental Protocol

The experimental methodology employs a sophisticated four-phase protocol designed to evaluate baseline model performance, the effectiveness of dual preference-based optimization approaches, and comparative assessment of DPO methodologies. This protocol enables systematic assessment of model improvement through both synthetic and hybrid preference optimization while maintaining rigorous experimental controls and enabling comparative analysis between optimization strategies.

Phase 1: Baseline Evaluation with Pre-trained Models

Phase 1 establishes comprehensive baseline performance measurements using pre-trained models in their original configurations. This phase employs the complete multi-modal checklist generation framework, evaluating each model across all three operational modes (Full-Prompt, Extract-Only, and Hybrid) to establish performance baselines for subsequent comparison analysis.

The baseline evaluation protocol implements systematic assessment across all model-topic-mode combinations, creating a comprehensive performance matrix that captures baseline capabilities across the full experimental space. This phase provides the foundational data required for preference pair generation and enables quantification of improvement through subsequent optimization procedures.

Phase 2: DPO-Synthetic Training and Evaluation

Phase 2 implements the synthetic preference learning approach (DPO-Synthetic) through systematic training of models using preference pairs derived exclusively from Judge Agent evaluations. This phase employs rank-1 emails as preferred examples and systematically selected lower-performing outputs as rejected examples, creating comprehensive synthetic preference datasets for each model category.

The DPO-Synthetic training protocol implements standardized procedures across all model variants (M0001-M0005) while maintaining consistency in hyperparameter configurations, training duration, and convergence criteria. Following training completion, models undergo comprehensive evaluation using identical assessment protocols employed in Phase 1, enabling direct comparison of pre-training and post-training performance through systematic quantification of improvement metrics.

Phase 3: DPO-Hybrid Training and Evaluation

Phase 3 implements the hybrid human-synthetic preference learning approach (DPO-Hybrid) through strategic integration of human-authored email examples with synthetic preference data. This phase incorporates the 25 validated human-written emails corresponding to topics T0001-T0025 as preferred examples, while maintaining synthetic data generation procedures for topics T0026-T0100.

The DPO-Hybrid training protocol employs identical training procedures and hyperparameter configurations used in Phase 2, ensuring that methodological differences can be attributed to preference data composition rather than training procedure variations. Post-training evaluation follows standardized assessment protocols, generating performance measurements directly comparable to both baseline (Phase 1) and DPO-Synthetic (Phase 2) results.

Phase 4: Comparative Pipeline Re-deployment and Analysis

Phase 4 implements a novel comparative evaluation approach that deploys both DPO-optimized model variants back into the original three-agent pipeline for comprehensive performance assessment. This phase represents a critical methodological innovation that enables evaluation

of fine-tuning effectiveness within the complete system context rather than isolated model assessment.

The pipeline re-deployment protocol implements systematic evaluation procedures where both DPO-Synthetic and DPO-Hybrid optimized models generate new email content across all 25 topics using identical prompts and generation parameters employed in baseline evaluation. The Judge Agent applies consistent evaluation frameworks to assess improvement magnitude and comparative effectiveness between the two DPO approaches.

This comparative deployment methodology enables direct assessment of which DPO method produces superior performance improvements within the complete multi-agent evaluation framework. The approach provides practical insights into optimization effectiveness while maintaining ecological validity through evaluation within the intended deployment context.

Mode-Based Analysis Framework

The multi-phase protocol incorporates systematic comparison across the three checklist generation modes, enabling assessment of optimization effectiveness under different operational conditions across both DPO methods. This mode-based analysis framework tests whether DPO improvements are consistent across different analytical approaches, whether optimization benefits are mode-dependent, and how the two preference learning methods perform under varying operational constraints.

The framework enables evaluation of interaction effects between dual optimization approaches and operational modes, providing insights into the generalizability of preference-based improvements across different evaluation contexts. This analysis approach supports both aggregate performance assessment and fine-grained investigation of optimization effectiveness while enabling direct comparison of synthetic versus hybrid preference learning approaches across all operational modes.

3.6.2 Multi-Topic Comparative Framework

The comparative framework employs a comprehensive factorial design approach where each model generates content for every topic within the experimental dataset across all four experimental phases, creating an extensive matrix of model-topic-phase combinations for analysis. This approach ensures that performance assessments capture model-specific capabilities, topic-dependent variations, dual-method optimization effectiveness, and comparative performance assessment across diverse contexts.

Controlled variables within the experimental design include prompt standardization across all model-topic combinations, consistent input formatting and parameter specifications, uniform evaluation criteria application regardless of generating model, and standardized environmental conditions for model inference. These controls ensure that observed performance differences reflect genuine model capabilities rather than experimental artifacts.

Randomization procedures minimize potential bias through several mechanisms: random ordering of topic presentation to each model prevents sequential effects, randomized model evaluation order eliminates potential carry-over effects, and random sampling of evaluation criteria prioritization reduces systematic bias in assessment frameworks.

Performance Delta Methodology

The performance delta methodology provides quantitative assessment of fine-tuning effectiveness through systematic comparison of pre-training and post-training performance measurements. This methodology enables precise quantification of improvement magnitude while accounting for baseline performance variations across different models and contexts.

Performance delta calculations employ normalized improvement metrics that account for baseline performance levels, enabling fair comparison across models with different starting capabilities. The methodology incorporates statistical significance testing to distinguish genuine improvements from random variation, ensuring robust conclusions regarding optimization effectiveness.

The delta analysis framework supports both aggregate improvement assessment and fine-grained analysis of improvement patterns across different experimental conditions. This approach enables identification of optimal optimization strategies and assessment of improvement consistency across diverse evaluation contexts.

Figure 3.4: *Multi-Phase Experimental Design Workflow showing baseline evaluation, dual-method DPO training, and comparative pipeline re-deployment phases*

Table 3.5: *Multi-Phase Experimental Conditions Matrix: Phase × Mode × Model Combinations*

Phase	Mode	Model Category	Model Count	Total Conditions
Phase 1: Baseline	Full-Prompt	Small/Medium/Large	7	21
Phase 1: Baseline	Extract-Only	Small/Medium/Large	7	21
Phase 1: Baseline	Hybrid	Small/Medium/Large	7	21
Phase 2: DPO-Synthetic	Full-Prompt	Small/Medium	5	15
Phase 2: DPO-Synthetic	Extract-Only	Small/Medium	5	15
Phase 2: DPO-Synthetic	Hybrid	Small/Medium	5	15
Phase 3: DPO-Hybrid	Full-Prompt	Small/Medium	5	15
Phase 3: DPO-Hybrid	Extract-Only	Small/Medium	5	15
Phase 3: DPO-Hybrid	Hybrid	Small/Medium	5	15
Phase 4: Pipeline Comparison	All Modes	Small/Medium	5	15
Total Experimental Conditions				**153**

3.6.3 Consistency Sampling Methodology

A critical innovation in this research is the implementation of consistency sampling through multiple generation approach, where each model generates three independent responses for every topic. This methodology enables assessment of both average performance and consistency reliability across repeated generations, providing insights into model stability and predictability.

The triple-generation approach serves multiple analytical purposes: quantification of within-model variance across identical input conditions, identification of models with high

consistency versus those with variable output quality, assessment of optimal generation strategies for practical deployment scenarios, and establishment of confidence intervals for performance measurements.

Figure 3.5: *Enhanced experimental design workflow showing multi-modal framework, two-phase protocol, and consistency sampling*

Cross-validation strategies enhance reliability assessment through systematic rotation of evaluation procedures and independent validation of assessment criteria across different model-topic combinations. This approach ensures that evaluation frameworks maintain validity across the diverse range of content generated throughout the experimental process.

3.7 Evaluation Framework

The evaluation framework implements a novel multi-stage assessment methodology designed to provide comprehensive and objective analysis of generated email quality Bohnet et al. (2022), Pimentel et al. (2024). This framework combines automated evaluation procedures with systematic criteria development to ensure consistent and reliable performance measurement across all experimental conditions.

3.7.1 Binary Checklist Generation Methodology

The checklist generation methodology employs the Checklist Creator Agent to develop structured evaluation frameworks tailored to each generated email while maintaining consistency in assessment standards. Each checklist comprises binary evaluation criteria that address key aspects of email effectiveness: content relevance and accuracy, persuasive appeal and emotional engagement, structural coherence and organization, audience appropriateness and tone, and call-to-action clarity and effectiveness.

The binary nature of evaluation criteria eliminates subjective scoring ambiguity while enabling systematic aggregation of assessment results across multiple evaluation dimensions. Each criterion receives binary classification (pass/fail) with associated priority weighting to reflect relative importance within the overall assessment framework.

Priority weighting system development accounts for the varying significance of different evaluation criteria within fundraising email effectiveness. High-priority criteria include factual accuracy, ethical appropriateness, and clear charitable mission alignment. Medium-priority criteria encompass persuasive effectiveness, emotional appeal, and structural organization. Low-priority criteria address stylistic preferences and minor formatting considerations.

3.7.2 Judge Agent Evaluation Protocol

The Judge Agent implements an advanced systematic evaluation protocol that leverages GPT O3 Mini's enhanced reasoning capabilities to apply generated checklists consistently across all email samples while accounting for priority weighting in final scoring calculations Marjanović et al. (2025), Sui et al. (2025). The evaluation process follows a standardized sequence enhanced by multi-dimensional analytical capabilities: comprehensive checklist application

with binary assessment for each criterion, advanced reasoning-based consistency verification, priority-weighted scoring aggregation to produce overall quality measures, comparative ranking generation across model outputs for identical topics, and comprehensive consistency analysis across multiple generations from the same model.

Multi-Dimensional Scoring System

The multi-dimensional scoring system represents a methodological advancement that incorporates reasoning-based analysis capabilities into systematic evaluation procedures. The system employs GPT O3 Mini's advanced analytical capabilities to provide enhanced evaluation reliability through sophisticated reasoning processes, improved consistency in scoring decisions, and systematic bias reduction in comparative assessments.

The scoring system implements multiple evaluation dimensions that capture different aspects of email quality: content quality assessment through factual accuracy and relevance analysis, persuasive effectiveness evaluation through rhetorical structure analysis, audience appropriateness assessment through tone and messaging alignment, and technical quality evaluation through structural and formatting analysis.

The probability-based scoring methodology converts binary assessments into quantitative measures suitable for advanced statistical analysis. The enhanced scoring algorithm incorporates reasoning-based confidence measures, weights individual criteria according to established priority levels and context-specific importance, and aggregates results to produce normalized performance scores ranging from 0 to 100 for comparative analysis purposes.

Consistency Measurement Protocols

Advanced consistency measurement protocols enable systematic assessment of evaluation reliability across multiple dimensions and experimental conditions. These protocols implement cross-mode reliability assessment to evaluate consistency of evaluation outcomes across the three operational modes, temporal stability analysis to assess evaluation consistency across repeated applications, and inter-model reliability verification to ensure evaluation fairness across different model categories.

The consistency measurement framework employs statistical reliability measures that quantify evaluation stability across different conditions while identifying potential sources of evaluation variance. This approach enables systematic improvement of evaluation procedures and validation of assessment framework reliability.

Statistical Significance Testing

The evaluation framework incorporates comprehensive statistical significance testing procedures specifically designed for pre-training and post-training performance comparisons. These procedures employ appropriate statistical methods for paired comparison analysis, account for multiple testing corrections in comprehensive model comparisons, and implement effect size calculations to assess practical significance beyond statistical significance.

Statistical testing protocols enable robust conclusion formation regarding DPO effectiveness while accounting for baseline performance variations and experimental design complexity. The testing framework supports both aggregate performance assessment and fine-grained analysis of improvement patterns across different experimental conditions.

3.7.3 Multi-Stage Evaluation Protocol

The multi-stage evaluation protocol represents a comprehensive methodological framework designed to assess dual-method DPO effectiveness across multiple evaluation dimensions while maintaining consistency and reliability throughout the four-phase experimental design. This protocol extends traditional evaluation approaches to accommodate the complexity of comparative preference learning assessment while ensuring robust statistical analysis across all experimental conditions **??**.

Training Effectiveness Assessment

Training effectiveness assessment provides systematic evaluation of DPO optimization success through comprehensive analysis of preference pair accuracy, convergence characteristics, and learning progression metrics. The assessment begins with preference pair validation accuracy measurement that quantifies how effectively each DPO method learns to distinguish between preferred and rejected examples during training procedures.

The methodology implements systematic tracking of training convergence metrics including loss function progression, gradient norm stability, validation performance improvement, and convergence time analysis. These metrics enable comparative assessment of training efficiency between synthetic and hybrid preference learning approaches while identifying potential optimization challenges or advantages specific to each method.

Performance delta quantification measures the magnitude of improvement achieved through each DPO method using normalized improvement metrics that account for baseline performance variations across different models and topics. The assessment employs statistical significance testing to distinguish genuine optimization improvements from random variation while providing confidence intervals for improvement magnitude estimates.

Quality retention analysis ensures that DPO optimization maintains or improves evaluation consistency across multiple generations while avoiding performance degradation in non-target evaluation dimensions. This analysis addresses potential concerns regarding optimization overfitting or unintended quality reduction in specific evaluation criteria.

Pipeline Integration Evaluation

Pipeline integration evaluation assesses how effectively DPO-optimized models function within the complete three-agent evaluation framework while maintaining system coherence and evaluation validity. The evaluation begins with integration compatibility assessment that verifies optimized models maintain technical compatibility with existing system infrastructure including prompt formatting, generation parameters, and output specifications.

System performance consistency evaluation ensures that DPO improvements translate into measurable system-level performance gains rather than merely reflecting isolated model improvements. The assessment employs end-to-end evaluation procedures that measure complete pipeline performance from prompt input through final Judge Agent scoring to capture system-level optimization effectiveness.

Evaluation framework stability analysis verifies that Judge Agent assessment procedures remain consistent and reliable when evaluating DPO-optimized model outputs compared to baseline assessments. This analysis addresses potential evaluation bias that might arise from

systematic differences in optimized model output characteristics while ensuring comparative assessment validity.

Cross-agent interaction assessment evaluates whether DPO optimization affects the quality of Checklist Creator Agent evaluation criteria generation or Judge Agent scoring consistency, thereby ensuring that optimization benefits reflect genuine content quality improvements rather than evaluation framework artifacts.

Cross-Method Comparative Analysis

Cross-method comparative analysis enables systematic assessment of relative effectiveness between synthetic and hybrid preference learning approaches through rigorous statistical comparison and practical significance evaluation. The analysis implements paired comparison procedures that directly assess Method 1 versus Method 2 performance across identical evaluation conditions while controlling for baseline performance variations and topic-specific effects.

The methodology employs multi-dimensional performance comparison that simultaneously assesses both methods across all evaluation criteria categories including content quality, persuasive effectiveness, audience appropriateness, and technical quality. This comprehensive approach enables identification of method-specific strengths and limitations while providing insights into optimal application contexts for each approach.

Statistical robustness verification ensures that observed method differences exceed random variation through comprehensive significance testing, effect size quantification, and power analysis procedures. The analysis incorporates multiple comparison corrections to maintain statistical validity while providing practical significance thresholds for method selection decisions.

Generalization assessment evaluates whether method differences remain consistent across different model categories, topic types, and operational modes, thereby informing practical deployment recommendations and identifying contexts where specific methods demonstrate superior effectiveness.

Cross-Validation Between Training Effectiveness and Pipeline Performance

The methodology implements comprehensive cross-validation procedures that systematically verify the relationship between DPO training effectiveness measures and subsequent pipeline performance improvements. This cross-validation approach addresses fundamental questions regarding whether training metrics accurately predict real-world system improvement while ensuring robust correlation between optimization indicators and practical deployment outcomes.

Cross-validation procedures employ systematic comparison of training convergence metrics with post-deployment performance measurements to establish predictive validity of training indicators. The analysis includes correlation assessment between training loss reduction and Judge Agent score improvements, validation of preference accuracy metrics against pipeline evaluation outcomes, and systematic documentation of training-performance relationships across different model categories and DPO methods.

The framework implements statistical validation of training-deployment consistency through temporal correlation analysis that examines whether models demonstrating superior training

convergence characteristics consistently achieve better performance in pipeline re-deployment scenarios. This validation enables reliable prediction of deployment success based on training metrics while informing optimal training termination criteria and resource allocation decisions **?**.

Pipeline consistency verification ensures that training improvements translate reliably into system-level performance gains through comprehensive testing of model integration procedures, evaluation framework stability, and assessment consistency across multiple deployment cycles. This verification process prevents technical artifacts from confounding the relationship between training effectiveness and practical system improvement while ensuring robust translation of optimization benefits into operational contexts.

Meta-Evaluation Framework

The meta-evaluation framework provides systematic assessment of evaluation framework consistency and reliability across the multi-phase experimental design while identifying potential sources of evaluation bias or systematic error **?**. Meta-evaluation procedures begin with Judge Agent consistency analysis that quantifies evaluation reliability across different experimental phases and optimization conditions.

Cross-phase evaluation stability assessment ensures that Judge Agent scoring criteria and assessment procedures remain consistent throughout the experimental timeline while identifying any systematic drift in evaluation standards that might confound comparative analysis. The assessment employs inter-rater reliability measures adapted for automated evaluation contexts to quantify assessment consistency.

Evaluation validity verification confirms that observed performance improvements reflect genuine quality enhancements rather than evaluation artifacts or systematic bias through correlation analysis with independent quality measures and expert validation procedures. This verification process strengthens confidence in experimental conclusions while identifying potential limitations in evaluation approach.

The meta-evaluation framework implements systematic documentation of evaluation uncertainty and confidence measures that inform result interpretation and practical application recommendations. These measures enable transparent reporting of evaluation limitations while providing guidance for future research and practical deployment decisions.

Table 3.6: *Enhanced evaluation criteria categories and priority weighting structure*

Category	**Criteria**	**Priority**	**Weight**
Content Quality	Factual Accuracy	High	0.25
Content Quality	Relevance Analysis	High	0.20
Persuasive Effectiveness	Rhetorical Structure	Medium	0.15
Persuasive Effectiveness	Emotional Appeal	Medium	0.15
Audience Appropriateness	Tone Alignment	Medium	0.10
Technical Quality	Structure & Format	Low	0.15

Inter-model comparison metrics enable systematic assessment of relative performance across different model architectures and sizes. These metrics include absolute performance scores for individual model-topic combinations, relative ranking positions within topic-specific

Table 3.7: *Evaluation Metrics and Statistical Tests by Analysis Type*

Analysis Type	Metrics	Statistical Test	Significance Level
Mode Comparison	Quality Score, Efficiency	ANOVA + Tukey HSD	p ¡ 0.05
Phase Comparison	Performance Delta	Paired t-test	p ¡ 0.01
Model Category	Aggregate Performance	Kruskal-Wallis	p ¡ 0.05
Consistency Analysis	Variance, CV	F-test	p ¡ 0.05
DPO Effectiveness	Improvement Ratio	Wilcoxon Signed-Rank	p ¡ 0.01
CrossMode Reliability	Cronbach's Alpha	Reliability Analysis	0.80

comparisons, consistency measures reflecting variance across multiple generations, and categorical performance analysis across small, medium, and large model groups.

3.8 Quality Assurance and Reliability

Quality assurance procedures ensure the integrity and reliability of experimental results through comprehensive validation mechanisms applied throughout the data collection and analysis processes. These enhanced procedures address potential sources of error, bias, and inconsistency that could compromise the validity of research findings, while incorporating specialized validation protocols for multi-modal operations and DPO training effectiveness.

3.8.1 Consistency Measurement and Validation

Consistency measurement across multiple generations provides crucial insights into model reliability and predictability. The measurement framework quantifies variation through statistical analysis of performance differences across the three generations per model-topic combination. Consistency metrics include standard deviation of performance scores across generations, coefficient of variation to normalize consistency measures across different performance levels, and range analysis to identify maximum performance variation within model outputs.

Output validation mechanisms verify the structural and content integrity of generated emails through automated checking procedures. Validation criteria include proper email formatting compliance, content length within specified parameters, topic relevance verification through keyword analysis, and ethical content screening to ensure appropriate charitable representation.

Bias identification and mitigation strategies address potential systematic influences on experimental results. Bias assessment includes analysis of model-specific performance patterns that might reflect training data characteristics, evaluation criteria bias that might favor particular model architectures or approaches, and temporal bias from sequential processing that might influence generation quality.

3.8.2 Mode Validation Procedures

Mode validation procedures ensure that each operational mode within the multi-modal checklist generation framework produces valid and reliable evaluation criteria while maintaining consistency with the overall evaluation objectives. These procedures implement systematic

validation protocols designed to verify mode-specific functionality and ensure that each mode achieves its intended methodological purpose.

Full-Prompt Mode Validation

Full-Prompt mode validation verifies that comprehensive contextual analysis produces high-quality evaluation criteria that capture all relevant aspects of email effectiveness. Validation procedures include assessment of criteria comprehensiveness to ensure coverage of all evaluation dimensions, verification of contextual relevance to confirm that criteria reflect specific email content and topic characteristics, and quality threshold analysis to establish that criteria meet minimum standards for evaluation effectiveness.

Extract-Only Mode Validation

Extract-Only mode validation confirms that minimal context processing maintains acceptable evaluation quality while achieving the intended efficiency improvements. Validation protocols include efficiency measurement verification to confirm processing time reductions, quality threshold maintenance to ensure evaluation criteria meet minimum effectiveness standards, and comparative analysis with Full-Prompt mode to quantify the quality-efficiency trade-off.

Hybrid Mode Validation

Hybrid mode validation ensures that two-step systematic analysis achieves the intended balance between comprehensive evaluation and processing efficiency. Validation procedures include systematic verification of both extraction and processing phases, assessment of phase integration effectiveness to ensure coherent evaluation criteria generation, and comparative analysis with both Full-Prompt and Extract-Only modes to confirm intermediate positioning in the quality-efficiency spectrum.

3.8.3 Cross-Mode Consistency Verification

Cross-mode consistency verification implements systematic procedures to validate evaluation reliability and fairness across all three operational modes. These procedures ensure that mode-specific differences reflect intended methodological variations rather than systematic bias or evaluation artifacts.

Inter-Mode Reliability Assessment

Inter-mode reliability assessment employs statistical correlation analysis to evaluate consistency of evaluation outcomes across different operational modes. Assessment procedures include calculation of inter-mode correlation coefficients for evaluation scores, analysis of ranking consistency across modes for identical model-topic combinations, and identification of systematic bias patterns that might indicate mode-specific evaluation artifacts.

Cross-Mode Fairness Verification

Cross-mode fairness verification ensures that evaluation procedures maintain fairness across different models and topics regardless of operational mode. Verification procedures include

assessment of mode-specific performance patterns to identify potential bias, analysis of score distribution characteristics across modes to ensure statistical comparability, and systematic evaluation of whether mode differences reflect genuine methodological variations rather than evaluation artifacts.

3.8.4 DPO Training Validation

DPO training validation procedures verify successful fine-tuning outcomes while ensuring that optimization improvements reflect genuine performance enhancements rather than training artifacts. These procedures implement comprehensive assessment of training effectiveness and model improvement validation.

Training Convergence Verification

Training convergence verification employs systematic monitoring of training metrics to ensure that DPO procedures achieve stable optimization outcomes. Verification procedures include loss function monitoring to confirm convergence achievement, gradient norm analysis to verify training stability, and validation performance tracking to ensure consistent improvement without overfitting.

Optimization Effectiveness Validation

Optimization effectiveness validation confirms that post-training performance improvements represent genuine enhancements in email generation quality. Validation procedures include comparative performance analysis to verify improvement magnitude, statistical significance testing to confirm that improvements exceed random variation, and systematic assessment of improvement consistency across different evaluation contexts.

3.8.5 Multi-Method Validation Framework

The multi-method validation framework implements comprehensive quality assurance procedures specifically designed for dual-method DPO evaluation while ensuring methodological rigor across all experimental phases. This framework addresses the unique challenges of validating human-synthetic data integration, pipeline consistency, and evaluation reliability in the context of comparative preference learning assessment **??**.

Human Data Quality Validation Protocols

Human data quality validation protocols ensure that human-authored email samples meet rigorous quality standards while maintaining compatibility with the synthetic evaluation framework. The validation process begins with expert review procedures involving fundraising professionals and communication specialists who assess email quality, effectiveness, and representativeness of professional fundraising communications.

Content authenticity verification confirms that human examples represent genuine best practices in fundraising communication through systematic analysis of persuasive effectiveness, structural coherence, audience appropriateness, and ethical compliance. The verification process includes linguistic analysis to confirm professional communication standards,

contextual relevance assessment to ensure topic alignment, and comparative quality analysis to verify that human examples represent meaningful quality improvements over synthetic alternatives.

Integration compatibility assessment ensures that human-authored emails maintain technical and methodological compatibility with the synthetic evaluation framework. Assessment procedures include formatting standardization verification, evaluation criteria applicability testing, scoring methodology compatibility analysis, and statistical distribution analysis to confirm that human examples do not introduce systematic bias into the preference learning framework.

Synthetic-Human Data Integration Verification

Synthetic-human data integration verification implements systematic procedures to ensure seamless combination of human and synthetic preference data while preserving the distinct characteristics and advantages of each data source. The verification process employs statistical analysis to confirm balanced representation across charitable categories, quality threshold consistency between data sources, and appropriate integration ratios for optimal preference learning effectiveness.

Distribution analysis procedures verify that human-synthetic integration maintains statistical validity through assessment of quality score distributions, evaluation criteria coverage, topic representation balance, and preference pair quality consistency. These analyses ensure that hybrid preference learning achieves intended benefits while avoiding systematic bias or integration artifacts that might compromise training effectiveness.

Cross-validation procedures verify integration success through independent assessment of preference pair quality using alternative evaluation frameworks and expert validation protocols. This multi-perspective validation approach strengthens confidence in integration effectiveness while identifying potential limitations or improvement opportunities in the hybrid approach.

Pipeline Consistency Across Fine-tuned Model Deployment

Pipeline consistency verification ensures that DPO-optimized models integrate seamlessly into the three-agent evaluation framework while maintaining system coherence and assessment reliability. The verification process implements comprehensive testing of model loading procedures, inference consistency, output formatting compliance, and evaluation framework compatibility to prevent technical artifacts from affecting comparative assessment.

System-level validation procedures confirm that fine-tuned models maintain compatibility with existing pipeline infrastructure through systematic testing of input/output interfaces, generation parameter consistency, prompt formatting requirements, and evaluation criteria applicability. These procedures ensure that observed performance differences reflect genuine optimization benefits rather than integration issues or technical incompatibilities.

Environmental consistency monitoring maintains standardized computational conditions across all evaluation phases through systematic documentation of hardware configurations, software dependencies, inference parameters, and evaluation timing. This monitoring ensures that comparative assessments accurately reflect model improvement rather than environmental variations or technical confounds.

Judge Agent Reliability Assessment Across Multiple Evaluation Cycles

Judge Agent reliability assessment implements systematic procedures to monitor evaluation consistency and identify potential sources of bias or drift that might develop through repeated application across multiple experimental phases. The assessment employs statistical analysis of evaluation patterns, scoring consistency, and criteria application to ensure reliable and fair assessment throughout the experimental timeline.

Temporal stability analysis evaluates whether Judge Agent assessment procedures remain consistent across the four experimental phases through correlation analysis of evaluation scores, consistency measurement of criteria application, and identification of systematic changes that might affect comparative validity. This analysis enables early detection of evaluation drift while ensuring robust comparative assessment.

Inter-phase reliability verification confirms that evaluation standards remain stable throughout the experimental process through systematic comparison of evaluation outcomes across phases, assessment of scoring distribution characteristics, and identification of potential systematic bias sources. The verification process includes statistical testing of evaluation consistency and confidence interval analysis for assessment reliability.

Cross-validation with alternative evaluation frameworks provides independent verification of Judge Agent assessment quality through comparison with expert evaluation protocols, alternative automated assessment approaches, and human baseline comparisons. This multi-method validation strengthens confidence in evaluation reliability while identifying potential limitations in the automated assessment framework.

3.8.6 Reproducibility and Documentation Standards

Reproducibility measures ensure that experimental procedures can be replicated by independent researchers with access to the same models and datasets while meeting contemporary standards for transparent and accountable research in language model evaluation **?**. Documentation standards include comprehensive recording of model configurations and parameters, detailed prompt specifications and input formatting procedures, complete evaluation criteria definitions and weighting schemes, and statistical analysis procedures with software version specifications.

The methodology implements standardized documentation protocols that address the unique challenges of dual-method DPO evaluation including preference data composition documentation, human-synthetic integration procedures, pipeline deployment specifications, and comparative evaluation protocols. These protocols ensure that all aspects of the experimental approach can be independently replicated while maintaining transparency in methodological choices and analytical procedures.

Data integrity verification protocols monitor the experimental process to identify and correct potential data collection errors while ensuring compliance with reproducibility standards. Verification procedures include automated checking of complete data collection across all model-topic combinations, validation of evaluation scoring calculations and aggregation procedures, cross-reference verification between generated content and corresponding evaluation results, and systematic documentation of any issues or deviations encountered during experimental execution.

Figure 3.6: *Quality Assurance Validation Framework including mode validation, cross-mode consistency, and DPO training verification*

Figure 3.7: *Mode Performance Comparison Visualization showing quality-efficiency trade-offs across operational modes*

3.9 Iterative Evaluation Methodology

The iterative evaluation methodology implements a systematic framework for conducting multi-phase assessment while maintaining consistency and reliability across the complete experimental timeline. This methodology addresses the unique challenges of evaluating dual-method DPO effectiveness through systematic procedures that ensure fair comparison while accounting for potential temporal effects and evaluation artifacts **??**.

3.9.1 Pipeline Re-deployment Framework

The pipeline re-deployment framework provides systematic procedures for integrating fine-tuned models into the operational three-agent evaluation system while maintaining experimental validity and assessment consistency. This framework represents a critical methodological innovation that enables evaluation of optimization effectiveness within realistic deployment contexts rather than isolated model assessment.

Model Integration Procedures for Fine-tuned Variants

Model integration procedures implement systematic protocols for replacing baseline models with their corresponding DPO-optimized variants within the Email Generation Agent while preserving system integrity and evaluation consistency. The integration process begins with comprehensive compatibility verification that ensures fine-tuned models maintain technical compatibility with existing system infrastructure including input/output formatting, parameter specifications, and interface requirements.

The methodology employs standardized substitution protocols that minimize technical artifacts while ensuring that observed performance differences reflect genuine optimization benefits rather than integration issues. Integration verification includes systematic testing of model loading procedures, prompt formatting consistency, generation parameter compatibility, and output validation to prevent technical confounds from affecting comparative assessment.

Version control and documentation procedures ensure complete traceability of model variants throughout the experimental process while enabling systematic comparison of integration success across different optimization methods. The framework implements automated verification procedures that confirm successful integration while identifying any technical issues that might compromise evaluation validity.

Consistency Maintenance Across Evaluation Phases

Consistency maintenance procedures ensure that evaluation standards and assessment criteria remain stable throughout the multi-phase experimental timeline while preventing systematic drift in evaluation quality or criteria application. The methodology implements systematic monitoring of Judge Agent evaluation consistency through statistical analysis of scoring patterns, assessment criteria application, and evaluation reliability measures across different experimental phases.

Temporal stability assessment employs longitudinal analysis of evaluation patterns to identify potential systematic changes in assessment standards that might confound comparative analysis between experimental phases. The assessment includes correlation analysis of evaluation scores across phases, consistency measurement of criteria application, and identification of potential evaluation drift that might affect result interpretation.

Environmental control procedures maintain consistent computational and software environments across all evaluation phases to prevent technical variations from affecting assessment outcomes. Controls include standardized hardware configurations, consistent software dependencies, identical inference parameters, and synchronized evaluation timing to minimize external sources of variation.

Bias Assessment for Repeated Judge Agent Evaluation

Bias assessment procedures systematically evaluate potential sources of systematic error that might arise from repeated application of the Judge Agent evaluation framework across multiple experimental phases. The assessment addresses concerns regarding evaluation consistency, potential adaptation effects, and systematic bias that might develop through repeated exposure to similar evaluation tasks.

The methodology implements statistical analysis of evaluation patterns to identify systematic bias including analysis of score distribution characteristics across phases, assessment of potential evaluation drift over time, identification of systematic preferences that might develop through repeated evaluation, and correlation analysis to detect potential bias sources. These procedures ensure that observed performance differences reflect genuine optimization effects rather than evaluation artifacts.

Cross-validation procedures verify evaluation consistency through independent validation of assessment outcomes using alternative evaluation frameworks and expert assessment protocols. This validation approach strengthens confidence in evaluation reliability while identifying potential limitations in the automated assessment framework that might affect result interpretation.

3.9.2 Statistical Methodology for Comparing Multi-Stage Results

The statistical methodology for multi-stage comparison implements comprehensive analytical procedures designed to assess dual-method DPO effectiveness while accounting for the complexity of multi-phase experimental design and potential temporal effects. This methodology extends traditional statistical approaches to accommodate the unique challenges of comparative preference learning assessment.

Multi-Level Modeling for Nested Experimental Design

Multi-level modeling procedures account for the hierarchical structure of the experimental design including models nested within categories, topics nested within charitable domains, and evaluation phases nested within the overall experimental timeline. The modeling approach enables simultaneous assessment of multiple factors while preserving the statistical relationships between different levels of experimental organization.

The methodology employs mixed-effects modeling procedures that account for both fixed effects (experimental conditions) and random effects (model-specific, topic-specific, and temporal variations) while providing robust estimation of treatment effects and their associated uncertainty. This approach enables valid statistical inference while accounting for the complex correlation structure inherent in the multi-phase experimental design.

Effect size estimation procedures provide practical significance assessment beyond statistical significance testing through comprehensive calculation of standardized effect sizes, confidence intervals for effect magnitude, and power analysis for detecting meaningful differences between DPO methods. These procedures inform practical deployment decisions while ensuring robust statistical conclusions.

Longitudinal Analysis of Performance Patterns

Longitudinal analysis procedures systematically assess performance patterns across the four experimental phases while identifying systematic trends, temporal effects, and optimization trajectories for each DPO method. The analysis employs time-series analytical techniques adapted for experimental contexts to quantify performance changes and identify optimal timing for assessment procedures.

Trend analysis procedures evaluate whether performance improvements represent stable optimization benefits or temporary effects that might diminish over time through systematic assessment of performance stability, evaluation of improvement persistence, and identification of potential performance degradation or enhancement over the experimental timeline.

Comparative trajectory analysis enables direct assessment of optimization effectiveness between DPO methods through systematic comparison of improvement patterns, convergence characteristics, and performance stability across the complete experimental timeline. This analysis provides insights into relative optimization efficiency while informing practical deployment recommendations.

Figure 3.8: *Iterative Evaluation Methodology Framework showing multi-phase consistency maintenance and bias assessment procedures*

3.10 Data Collection Procedures

Data collection procedures implement a systematic pipeline designed to ensure comprehensive and consistent gathering of experimental data across all model-topic combinations while maintaining quality standards and enabling efficient analysis of results.

Table 3.8: *Iterative Evaluation Metrics and Consistency Measures*

Evaluation Aspect	Metric	Measurement	Threshold
Temporal Consistency	Correlation across phases	Pearson r	r ¿ 0.85
Evaluation Stability	Score variance	Coefficient of variation	CV ¡ 0.15
Bias Assessment	Systematic drift	Linear trend analysis	p ¿ 0.05
Integration Success	Technical compatibility	Error rate	¡ 2%
Phase Reliability	Inter-phase agreement	Cronbach's alpha	¿ 0.80

3.10.1 Systematic Generation Pipeline

The generation pipeline orchestrates the sequential application of all models to every topic within the experimental dataset, ensuring consistent conditions and comprehensive coverage of all required model-topic combinations. Pipeline implementation includes automated model loading and configuration management, standardized prompt application with consistent formatting across all models, systematic generation scheduling to optimize computational resource utilization, and automated storage of generated content with appropriate metadata and identification.

The pipeline incorporates error handling and recovery mechanisms to address potential model failures or generation issues without compromising experimental integrity. Recovery procedures include automatic retry mechanisms for failed generations, alternative generation strategies for models with specific configuration requirements, and comprehensive logging of any issues encountered during the generation process.

3.10.2 Automated Evaluation and Scoring

Automated evaluation procedures apply the three-agent assessment framework systematically across all generated content, ensuring consistent evaluation standards while minimizing manual intervention requirements. The automated scoring system includes sequential application of checklist generation and evaluation procedures, standardized scoring calculations with priority weighting, and systematic aggregation of results across multiple generations per model-topic combination.

Cross-model performance measurement protocols enable fair comparison across diverse model architectures and capabilities. Measurement standardization includes normalized scoring procedures that account for different model output characteristics, consistent evaluation timeframes to ensure equal assessment opportunity for all models, and systematic application of identical evaluation criteria regardless of generating model.

Result aggregation and storage methodology organizes experimental data for efficient analysis while preserving complete information for detailed investigation and validation. Storage procedures include structured database organization with comprehensive metadata, automated backup and version control for data integrity, and export capabilities for statistical analysis software compatibility.

Statistical analysis preparation procedures ensure that collected data meets the requirements for planned analytical approaches while identifying any data quality issues that might affect subsequent analysis. Preparation includes verification of complete data collection across all experimental conditions, assessment of data distribution characteristics for appropriate

Table 3.9: *Data collection metrics and completeness verification*

Collection Phase	Expected Items	Collected Items	Success Rate
[Data collection metrics table to be completed]			

statistical test selection, and identification of outliers or anomalous results requiring further investigation.

3.11 Direct Preference Optimization Implementation

The methodology incorporates a comprehensive Direct Preference Optimization (DPO) implementation that enhances model performance based on comparative evaluation results from the multi-agent evaluation framework. This implementation represents a significant methodological advancement that leverages systematic preference data generation to achieve targeted model improvement through theoretically grounded optimization procedures.

3.11.1 DPO Methodology Framework

Direct Preference Optimization provides a theoretically grounded approach to fine-tuning language models using preference data derived from comparative evaluations Rafailov et al. (2023), Muldrew et al. (2024). Unlike traditional reinforcement learning from human feedback (RLHF) approaches, DPO directly optimizes the policy model using preference pairs without requiring a separate reward model, offering computational efficiency and training stability advantages Wang et al. (2024).

Preference Pair Generation from Judge Agent Scores

The DPO methodology framework implemented in this research utilizes a sophisticated preference pair generation system based on comparative evaluation results from the Judge Agent. The system employs systematic analysis of performance scores across model-topic combinations to identify optimal preference pairs that capture meaningful quality distinctions while ensuring statistical robustness.

Preference pair selection follows rigorous criteria designed to maximize training effectiveness: score differential thresholds ensure substantial quality differences between preferred and non-preferred examples, topic consistency requirements maintain contextual coherence within preference pairs, and statistical significance verification confirms that observed performance differences exceed random variation. This systematic approach ensures that preference data directly reflects genuine quality distinctions identified through the multi-agent evaluation framework.

The preference pair generation algorithm implements automated quality controls that include minimum performance threshold requirements for preferred examples, maximum threshold restrictions for non-preferred examples to avoid training on fundamentally flawed outputs, balanced topic distribution to prevent domain-specific bias, and cross-validation procedures to verify preference pair quality across different evaluation dimensions.

Training Pipeline Methodology

The training pipeline implements a systematic approach to DPO fine-tuning that ensures reproducible and effective model optimization. The pipeline incorporates automated data preprocessing procedures that convert Judge Agent evaluations into properly formatted preference datasets, systematic hyperparameter optimization based on model size and architectural characteristics, and comprehensive training monitoring to ensure convergence and prevent overfitting.

Training data preparation follows a multi-stage validation process that begins with comprehensive analysis of baseline evaluation results to identify optimal preference pairs. The process includes statistical analysis of score distributions to establish appropriate threshold criteria, topic-balanced sampling to ensure representative coverage across charitable categories, and quality verification procedures to confirm preference pair validity.

The systematic training methodology employs standardized procedures across all model variants to ensure fair comparison of optimization effectiveness. Training protocols include consistent batch size selection based on model capacity and computational constraints, learning rate optimization through systematic hyperparameter search, and convergence monitoring through validation loss tracking and performance plateau detection.

3.11.2 Dual-Method DPO Experimental Design

This research introduces a novel dual-method approach to Direct Preference Optimization that systematically compares the effectiveness of purely synthetic preference data against hybrid human-synthetic preference learning. This methodological innovation addresses fundamental questions regarding the optimal composition of preference datasets while establishing empirical evidence for best practices in preference-based fine-tuning **??**.

DPO-Synthetic: Synthetic Preference Learning

DPO-Synthetic implements a fully synthetic approach to preference data generation, utilizing the highest-ranked emails from the Judge Agent evaluation framework as preferred examples in DPO training. This approach leverages the systematic evaluation capabilities of the multi-agent framework to create preference pairs based entirely on model-generated content and automated assessment procedures.

The synthetic preference learning methodology follows a systematic data curation process that begins with comprehensive analysis of Judge Agent scores across all model-topic combinations. Preferred examples are selected based on rank-1 performance within each topic category, ensuring that chosen responses represent the highest quality outputs as determined by the standardized evaluation framework. Rejected examples are systematically selected from lower-performing outputs within the same topic categories, maintaining contextual consistency while ensuring substantial quality differentials.

This methodology offers several theoretical advantages for preference optimization: scalability through automated data generation procedures, consistency in evaluation standards across all preference pairs, elimination of human annotation subjectivity and potential bias, and direct integration with the existing multi-agent evaluation framework. The approach enables comprehensive preference dataset generation across all 25 topics while maintaining systematic quality controls and statistical robustness **?**.

Quality assurance procedures for synthetic preference data include statistical validation of score differentials between preferred and rejected examples, cross-modal consistency verification across different checklist generation modes, topic-balanced distribution to prevent domain-specific bias, and automated filtering to remove low-confidence preference pairs based on evaluation score variance.

DPO-Hybrid: Hybrid Human-Synthetic Learning

DPO-Hybrid implements a hybrid approach that strategically integrates human-authored email examples with synthetic preference data to assess the potential benefits of human expertise in preference learning. This methodology incorporates 25 high-quality human-written fundraising emails corresponding to topics T0001-T0025, while maintaining synthetic data generation for the remaining 75 topics in the experimental dataset.

The hybrid integration methodology employs a systematic approach to combining human and synthetic preference data that ensures compatibility and maintains training effectiveness. Human-authored emails undergo comprehensive quality validation through expert review processes involving fundraising professionals and communication specialists. These emails serve as preferred examples within their respective topic categories, while rejected examples are generated through systematic selection of lower-performing model outputs for identical topics.

Theoretical justification for the hybrid approach stems from recent advances in preference learning research that demonstrate enhanced training effectiveness when human expertise is strategically integrated with synthetic data generation **?**. The approach addresses potential limitations of purely synthetic preference learning by incorporating authentic human expertise while maintaining the scalability benefits of automated data generation for the majority of the training dataset.

Human data integration procedures include rigorous quality validation protocols to ensure consistency with synthetic data standards, topic-specific alignment verification to maintain contextual coherence, statistical analysis to confirm that human examples represent genuine quality improvements over synthetic alternatives, and systematic documentation of human-synthetic integration ratios for subsequent analysis.

Comparative Theoretical Framework

The dual-method experimental design enables systematic investigation of fundamental questions in preference learning research: the relative effectiveness of purely synthetic versus human-grounded preference data, the optimal ratio of human to synthetic examples for training effectiveness, the generalizability of preference learning across different data composition strategies, and the practical implications of hybrid approaches for scalable preference optimization.

Theoretical expectations for the comparative analysis include enhanced performance from hybrid human-synthetic learning due to incorporation of authentic human expertise, improved generalization capabilities through exposure to diverse preference sources, potential efficiency gains from strategic human data integration rather than comprehensive human annotation, and insights into optimal data composition strategies for practical deployment scenarios.

The comparative framework implements rigorous experimental controls to ensure valid assessment of methodological differences. Both methods employ identical training procedures, hyperparameter configurations, and evaluation frameworks, with the sole variable being the composition of preference training data. This approach enables precise attribution of performance differences to preference data characteristics rather than training procedure variations.

3.11.3 Fine-tuning Experimental Design

The fine-tuning experimental design implements a comprehensive controlled approach that enables systematic assessment of DPO effectiveness across different model categories and operational modes. The design incorporates rigorous baseline preservation through comprehensive documentation of pre-fine-tuning performance characteristics, controlled fine-tuning procedures with standardized hyperparameters and training protocols, and systematic post-fine-tuning evaluation using identical assessment frameworks applied in the baseline experimental phase.

Pre/Post-Training Comparative Framework

The comparative framework employs a sophisticated experimental design that enables precise quantification of DPO effectiveness through systematic comparison of baseline and optimized model performance. The framework implements controlled comparison protocols that ensure fair assessment of optimization benefits while accounting for potential confounding variables.

Pre-training documentation procedures establish comprehensive baseline measurements that include detailed performance profiles across all evaluation dimensions, consistency measurements across multiple generations, computational efficiency metrics including inference time and resource utilization, and systematic documentation of generation characteristics and quality patterns.

Post-training evaluation protocols employ identical assessment procedures to enable direct comparison while incorporating additional measurements specific to optimization assessment. These procedures include comparative performance analysis across identical model-topic-mode combinations, optimization effectiveness quantification through normalized improvement metrics, and systematic assessment of whether performance improvements are consistent across different operational contexts.

Model selection for fine-tuning encompasses both small and medium-sized models (M0001-M0005) that demonstrate adequate baseline performance while remaining computationally feasible for fine-tuning procedures. This approach enables comprehensive assessment of fine-tuning effectiveness across different model scales without excessive computational requirements while providing results applicable to practical deployment scenarios.

Preference data curation implements comprehensive quality controls to ensure that training examples reflect genuine performance improvements rather than evaluation artifacts. Curation procedures include rigorous minimum performance threshold requirements for preferred examples, systematic verification of substantial quality differences between preferred and non-preferred pairs, balanced topic distribution to prevent over-representation of specific charitable categories, and cross-validation procedures to verify preference pair consistency across different evaluation modes.

Figure 3.9: *DPO Training Pipeline Architecture showing data flow from evaluation to training and validation*

Table 3.10: *DPO Training Configuration and Hyperparameters*

Parameter	Small Models	Medium Models	Justification
Batch Size	4	2	Memory optimization
Learning Rate	5e-6	3e-6	Stability & convergence
Training Epochs	3	2	Prevent overfitting
LoRA Rank	16	32	Parameter efficiency
LoRA Alpha	32	64	Adaptation strength
Dropout	0.1	0.1	Regularization
Beta (DPO)	0.1	0.1	Preference strength
Max Length	1024	1024	Context limitation
Warmup Steps	100	150	Learning rate scheduling

3.11.4 Post-Fine-tuning Evaluation Protocol

Post-fine-tuning evaluation employs identical assessment procedures used in the initial experimental phase to ensure valid comparison of pre- and post-fine-tuning performance. The evaluation protocol maintains consistency in topic selection, prompt formatting, generation parameters, and assessment criteria application while documenting any observed changes in generation characteristics or evaluation outcomes.

Comparative analysis framework enables systematic assessment of fine-tuning effectiveness through multiple evaluation dimensions: absolute performance improvement measurement across all evaluation criteria, relative performance changes within specific charitable categories, consistency improvement assessment through variance analysis across multiple generations, and efficiency analysis comparing pre- and post-fine-tuning inference characteristics.

3.11.5 Post-Training Pipeline Evaluation

The post-training pipeline evaluation represents a critical methodological innovation that extends beyond traditional isolated model assessment to evaluate DPO effectiveness within the complete three-agent system context. This approach addresses fundamental questions regarding the translation of training improvements into practical system performance while enabling direct comparison of dual-method DPO effectiveness **??**.

Pipeline Re-deployment Methodology

The pipeline re-deployment methodology implements systematic procedures for integrating fine-tuned models back into the operational three-agent framework while maintaining evaluation consistency and experimental validity. The methodology begins with systematic model substitution procedures that replace baseline models with their corresponding DPO-optimized variants within the Email Generation Agent while preserving all other system components and configurations.

Integration verification protocols ensure that fine-tuned models maintain compatibility with existing pipeline infrastructure through comprehensive testing of model loading procedures, prompt formatting consistency, generation parameter compatibility, and output formatting compliance. This verification process prevents technical artifacts from confounding evaluation results while ensuring that observed performance differences reflect genuine optimization benefits rather than integration issues.

The re-deployment protocol maintains strict environmental controls that include identical hardware configurations across all evaluation phases, consistent software dependencies and version specifications, standardized inference parameters and generation settings, and synchronized evaluation timing to minimize external variability. These controls ensure that comparative assessments accurately reflect model improvement rather than environmental differences **?**.

Comparative DPO Method Assessment

The comparative assessment framework enables direct evaluation of which DPO method produces superior performance improvements through systematic application of both optimized model variants to identical evaluation scenarios. The assessment protocol implements parallel evaluation procedures where both DPO-Synthetic and DPO-Hybrid optimized models generate email content across the complete 25-topic evaluation dataset using identical prompts, generation parameters, and environmental conditions.

Performance measurement consistency maintains evaluation validity through systematic application of the original Judge Agent evaluation framework to assess improvement magnitude and comparative effectiveness. The assessment employs identical checklist generation procedures, evaluation criteria application, and scoring methodologies used in baseline evaluation to ensure direct comparability of results across all experimental phases.

The comparative framework implements statistical procedures designed to distinguish genuine method differences from random variation through rigorous significance testing, effect size quantification, and confidence interval analysis. These procedures enable robust conclusions regarding the relative effectiveness of synthetic versus hybrid preference learning approaches while accounting for baseline performance variations and evaluation uncertainty.

Ecological Validity Assessment

Ecological validity assessment ensures that optimization improvements translate into meaningful performance gains within realistic deployment contexts rather than merely reflecting training dataset overfitting or evaluation artifacts. The assessment employs systematic evaluation of model performance across diverse topic categories, operational modes, and generation scenarios to verify that improvements generalize beyond training conditions.

The methodology implements cross-modal validation procedures that assess whether DPO improvements remain consistent across the three checklist generation modes (Full-Prompt, Extract-Only, and Hybrid), thereby testing the robustness of optimization benefits under varying operational constraints. This cross-modal assessment provides insights into the generalizability of preference learning improvements while identifying potential mode-specific optimization effects.

Real-world applicability assessment evaluates whether optimized models maintain improvement consistency across the complete range of charitable categories and fundraising scenarios represented within the experimental dataset. This assessment addresses practical deployment considerations by verifying that optimization benefits extend beyond the specific topics and contexts used for preference pair generation **?**.

Figure 3.10: *Post-Training Pipeline Evaluation Framework showing dual-method comparison within the complete three-agent system*

Table 3.11: *Pipeline Evaluation Metrics and Comparative Assessment Framework*

Evaluation Dimension	**Baseline**	**DPO-Synthetic**	**DPO-Hybrid**
Generation Quality	Judge Agent Score	Judge Agent Score	Judge Agent Score
Consistency Variance	Multi-generation SD	Multi-generation SD	Multi-generation SD
Cross-Modal Stability	Mode Correlation	Mode Correlation	Mode Correlation
Topic Generalization	Category Coverage	Category Coverage	Category Coverage
Ecological Validity	Real-world Alignment	Real-world Alignment	Real-world Alignment
Statistical Comparison	**Baseline vs DPO-S**	**Baseline vs DPO-H**	**DPO-S vs DPO-H**

3.12 Performance Analysis Methods

The performance analysis methodology employs comprehensive statistical approaches designed to extract meaningful insights from the multi-dimensional experimental data while accounting for the complex relationships between model characteristics, topic variations, and evaluation outcomes.

3.12.1 Statistical Analysis Framework

The statistical analysis framework implements a multi-layered approach that addresses different aspects of model performance assessment through appropriate analytical techniques. Primary analysis focuses on comparative performance assessment across model categories using analysis of variance (ANOVA) procedures to identify significant differences between small, medium, and large model groups while controlling for topic-specific variations.

Significance testing methodology incorporates multiple comparison corrections to address the increased Type I error risk associated with numerous model-topic comparisons. Bonferroni correction procedures ensure that family-wise error rates remain within acceptable bounds while maintaining sufficient statistical power for meaningful effect detection. Effect size calculations using Cohen's d and eta-squared measures provide practical significance assessment beyond statistical significance testing.

Multi-dimensional performance assessment recognizes that email generation quality encompasses multiple evaluation criteria with varying importance levels. The analysis employs multivariate analysis of variance (MANOVA) procedures to assess simultaneous differences

across multiple evaluation dimensions while preserving the relationships between different quality aspects.

Comparative Analysis Methods for DPO Evaluation

The comparative analysis methodology implements specialized statistical procedures designed to assess dual-method DPO effectiveness while accounting for the complex multi-phase experimental design and nested data structure. This methodology extends traditional statistical approaches to accommodate the unique challenges of preference learning comparison and multi-stage evaluation assessment.

Multi-level modeling procedures account for the hierarchical structure of the experimental data including models nested within size categories, topics nested within charitable domains, and evaluation phases nested within the overall experimental timeline. The modeling approach employs mixed-effects procedures that simultaneously assess fixed effects (DPO methods, model categories, operational modes) and random effects (model-specific, topic-specific, and temporal variations) while providing robust estimation of treatment effects and their associated confidence intervals.

Effect size calculations for DPO method comparison implement comprehensive assessment of practical significance beyond statistical significance testing. The methodology employs specialized effect size measures adapted for preference learning contexts including standardized mean differences between DPO methods, proportion of variance explained by method differences, and practical significance thresholds calibrated for email generation quality assessment. These measures enable informed decisions regarding optimal DPO approaches while providing transparent reporting of improvement magnitude.

Confidence intervals for practical significance assessment provide robust uncertainty quantification for DPO method effectiveness while enabling reliable deployment decisions. The methodology implements bootstrap procedures and Bayesian estimation techniques to generate confidence intervals that account for experimental design complexity and multiple sources of variation. These intervals inform practical deployment recommendations while ensuring transparent reporting of analytical uncertainty.

Power analysis for detecting meaningful differences between methods ensures that the experimental design provides adequate statistical power for reliable detection of practically significant DPO method differences. The analysis employs simulation-based procedures that account for the multi-phase design complexity while providing prospective power estimates for different effect sizes and sample configurations. This analysis validates the experimental design while informing interpretation of null results and negative findings.

Effect Size Predictions for Each DPO Method

The methodology incorporates systematic predictions of expected effect sizes for both DPO methods based on theoretical considerations and empirical evidence from preference learning research. These predictions establish empirically grounded expectations for optimization effectiveness while providing benchmarks for evaluating actual experimental outcomes against theoretical predictions.

DPO-Synthetic effect size predictions are calibrated based on observed performance differentials in baseline evaluation data and theoretical expectations for purely automated prefer-

ence optimization. Expected effect sizes for synthetic preference learning range from medium effects (Cohen's d = 0.5) for content quality improvements to large effects (Cohen's d = 0.8) for consistency measures, reflecting the systematic nature of automated preference pair generation and the direct integration with the multi-agent evaluation framework.

DPO-Hybrid effect size predictions incorporate enhanced expectations based on the integration of authentic human expertise with systematic synthetic data generation. Expected effect sizes for hybrid preference learning range from large effects (Cohen's d = 0.8) for content quality and persuasive effectiveness to very large effects (Cohen's d 1.0) for audience appropriateness measures, reflecting the potential benefits of incorporating human communication expertise into the preference learning framework.

Comparative effect size predictions between DPO methods anticipate medium to large effects (Cohen's d = 0.5-0.8) favoring DPO-Hybrid across most evaluation dimensions, with particularly pronounced advantages expected in persuasive effectiveness and audience appropriateness categories. These predictions are based on theoretical advantages of human expertise integration and empirical evidence from related preference learning research **??**.

Practical Significance Thresholds for Method Selection

The methodology establishes systematic practical significance thresholds that inform evidence-based decisions regarding optimal DPO method selection for different deployment scenarios. These thresholds extend beyond statistical significance to address practical considerations including deployment cost, implementation complexity, and operational requirements for real-world applications.

Primary practical significance thresholds are calibrated for Judge Agent score improvements, with minimum meaningful improvement defined as 5-point increases on the 100-point scale for individual model categories and 3-point aggregate improvements across all models. These thresholds reflect practical requirements for detectable quality improvements in email generation systems while accounting for measurement uncertainty and evaluation consistency considerations.

Secondary practical significance thresholds address operational considerations including training resource requirements, with cost-effectiveness thresholds established at 10

Method selection criteria integrate practical significance assessment with statistical evidence through multi-criteria decision frameworks that systematically weight quality improvements, resource requirements, implementation complexity, and deployment reliability. These frameworks enable transparent and reproducible decisions regarding optimal DPO approaches while accounting for diverse practical constraints and deployment scenarios **?**.

Threshold validation procedures employ sensitivity analysis to assess the robustness of method selection decisions across different threshold specifications while ensuring that practical significance criteria remain appropriate for diverse deployment contexts. This validation approach strengthens confidence in method selection recommendations while providing transparency regarding the impact of threshold choices on final deployment decisions.

Reliability Measures Across Iterative Evaluations

The methodology implements comprehensive reliability measurement procedures that systematically assess evaluation consistency and dependability across the complete multi-phase

experimental timeline. These reliability measures address fundamental questions regarding evaluation stability, assessment consistency, and measurement precision while ensuring robust statistical foundations for comparative analysis and practical deployment decisions.

Temporal reliability assessment employs test-retest methodologies adapted for automated evaluation contexts to quantify assessment consistency across repeated applications of the Judge Agent evaluation framework. The assessment includes systematic measurement of evaluation score stability across multiple evaluation cycles, correlation analysis of assessment outcomes across different time periods, and identification of systematic drift or bias that might develop through repeated evaluation exposure.

Inter-phase reliability verification implements systematic procedures to ensure evaluation consistency across the four experimental phases while accounting for potential systematic changes in evaluation patterns or criteria application. The verification process employs statistical techniques including intraclass correlation coefficients (ICC) to quantify evaluation reliability, Cronbach's alpha analysis to assess internal consistency of evaluation criteria, and systematic assessment of evaluation score distributions across phases to identify potential systematic bias sources.

Cross-modal reliability analysis evaluates assessment consistency across the three operational modes (Full-Prompt, Extract-Only, and Hybrid) to ensure that evaluation reliability remains stable regardless of checklist generation approach. This analysis addresses concerns regarding mode-specific evaluation artifacts while confirming that reliability measures remain consistent across different analytical frameworks and operational constraints.

Multi-model reliability assessment examines evaluation consistency across different model categories and architectural types to verify that assessment reliability does not vary systematically based on model characteristics or output patterns. The assessment employs variance component analysis to quantify the relative contributions of model-specific, evaluation-specific, and random sources of variation to overall assessment reliability while ensuring fair and consistent evaluation across diverse model architectures.

Statistical reliability thresholds establish minimum acceptable levels of assessment consistency required for valid comparative analysis and deployment decisions. Primary reliability thresholds include ICC values exceeding 0.80 for inter-phase consistency, Cronbach's alpha coefficients greater than 0.85 for evaluation criteria internal consistency, and test-retest correlations above 0.90 for temporal stability across repeated evaluations. These thresholds ensure robust statistical foundations for comparative analysis while providing confidence in evaluation framework reliability.

Reliability improvement procedures implement systematic approaches to enhance evaluation consistency when reliability measures fall below established thresholds. Improvement strategies include criteria refinement to address sources of evaluation inconsistency, training data enhancement to improve Judge Agent assessment quality, and systematic bias correction to address identified sources of evaluation drift or systematic error. These procedures ensure continuous improvement in evaluation reliability while maintaining experimental integrity throughout the research timeline **??**.

3.12.2 Correlation and Trend Analysis

Correlation analysis investigates relationships between model characteristics and performance outcomes to identify systematic patterns that might inform model selection and deployment

Table 3.12: *Enhanced Statistical Analysis Plan for Dual-Method DPO Evaluation*

Analysis Type	Method	Purpose	Signifi
Baseline Comparison	Mixed-effects ANOVA	Model category differences	p
DPO-Synthetic vs Baseline	Paired t-test	Synthetic preference effectiveness	p
DPO-Hybrid vs Baseline	Paired t-test	Hybrid preference effectiveness	p
DPO-Synthetic vs DPO-Hybrid	Paired t-test	Comparative DPO effectiveness	p
Multi-phase Analysis	Repeated measures ANOVA	Temporal stability	p
Cross-modal Consistency	Friedman test	Mode-dependent effects	p
Effect size (DPO methods)	Cohen's d	Practical significance	d
Pipeline Integration	MANOVA	System-level improvement	p
Power Analysis	Simulation-based	Design validation	Pow
Multi-level Modeling	Mixed-effects	Nested design	AIC/BI

decisions. The analysis examines correlations between parameter count and overall performance scores, architectural features and specific evaluation criteria performance, and consistency measures and model reliability characteristics.

Trend identification across model categories employs regression analysis techniques to quantify performance scaling relationships and identify optimal model size categories for different deployment scenarios. Polynomial regression models assess non-linear relationships between model size and performance while accounting for diminishing returns that might occur with increasing parameter counts.

Temporal stability analysis examines performance consistency across the multiple generations per model-topic combination to assess model reliability and predictability. Variance component analysis quantifies the relative contributions of model-specific, topic-specific, and random variation sources to overall performance variation, informing model selection criteria for practical applications.

3.13 Result Validation and Interpretation

Result validation procedures ensure the reliability and generalizability of research findings through comprehensive verification mechanisms that address potential sources of bias, error, and misinterpretation while establishing confidence in the research conclusions.

3.13.1 External Validation Procedures

External validation employs independent assessment mechanisms to verify the reliability of the automated evaluation framework and validate the meaningfulness of identified performance differences. Expert evaluation integration involves fundraising professionals and communication specialists in reviewing selected email samples to assess whether automated evaluation outcomes align with human judgment regarding email quality and effectiveness.

The expert evaluation protocol implements a structured approach that parallels the automated assessment framework while allowing for nuanced human judgment. Expert evaluators assess the same email samples evaluated by the Judge Agent using comparable criteria but with the flexibility to provide qualitative feedback and identify aspects of quality that might

not be captured through automated assessment procedures.

Human baseline comparison methodology establishes performance benchmarks through human-generated email samples for the same topics used in model evaluation. This comparison enables assessment of whether language model performance approaches or exceeds human-level quality in fundraising email generation while identifying specific areas where models demonstrate particular strengths or limitations.

Figure 3.11: *Multi-layer validation framework showing expert evaluation, human baseline comparison, and cross-validation procedures*

3.13.2 Interpretation Framework and Limitations

The result interpretation framework provides systematic guidelines for drawing valid conclusions from experimental data while acknowledging inherent limitations and potential alternative explanations for observed outcomes. Interpretation procedures include assessment of practical significance beyond statistical significance, consideration of confidence intervals and effect size magnitudes, evaluation of result consistency across different analytical approaches, and identification of potential confounding variables or alternative explanations.

Limitation identification addresses several key areas that might affect result generalizability: domain specificity considerations regarding the focus on charity fundraising emails, evaluation framework limitations including potential bias in automated assessment criteria, model selection constraints related to the focus on open-source implementations, and temporal considerations regarding the snapshot nature of model capabilities assessment.

Generalizability assessment examines the extent to which findings might apply to broader email generation tasks beyond the specific charitable fundraising domain. This assessment considers the transferability of evaluation methodologies to other persuasive communication contexts, the applicability of model performance patterns to different content domains, and the relevance of identified optimization strategies for broader automated content generation applications.

The comprehensive methodology presented in this chapter provides a robust framework for evaluating language model performance in automated email generation while establishing important precedents for multi-agent evaluation systems and preference-based fine-tuning approaches. The three-stage approach ensures systematic progression from foundational design through implementation to advanced analysis and enhancement, supporting both immediate research objectives and longer-term methodological contributions to the field.

Chapter 4

Results

4.1 Agent Model Selection Validation

Systematic comparison studies were conducted to evaluate the performance of traditional language models versus reasoning-capable models for checklist generation and evaluation tasks. For the Checklist Creator Agent, comparative analysis between GPT-4.1-nano (representing traditional high-performance models) and DeepSeek R1 (representing reasoning-specialized architectures) demonstrated significant advantages in evaluation criteria quality and consistency for the reasoning-capable model.

Similarly, for the Judge Agent, empirical comparison between Gemini-2.5 Flash (representing efficient traditional models) and GPT O3 Mini (representing reasoning-optimized models) revealed superior performance in evaluation consistency, scoring reliability, and analytical depth for the reasoning-capable architecture. These preliminary studies established the empirical foundation for agent-specific model selection based on task-appropriate capabilities rather than general performance metrics.

Table 4.1: *Comparative Performance of Model Types for Agent Tasks*

Agent	Traditional Model	Reasoning Model	Performance Metric	Improvement
Checklist Creator	GPT-4.1-nano	DeepSeek R1	Criteria Quality	+23%
Checklist Creator	GPT-4.1-nano	DeepSeek R1	Consistency Score	+18%
Judge Agent	Gemini-2.5 Flash	GPT O3 Mini	Evaluation Reliability	+31%
Judge Agent	Gemini-2.5 Flash	GPT O3 Mini	Scoring Consistency	+27%

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. Pha-

sellus 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 5

Discussion and Conclusions

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

Bohnet, B., Tran, V. Q., Verga, P., Aharoni, R., Andor, D., Soares, L. B., Ciaramita, M. & Eisenstein, J. (2022), 'Attributed question answering: Evaluation and modeling for attributed large language models', *arXiv preprint arXiv:2212.08037* .

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* .

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* .

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* .

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

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* .

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* .

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* .

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

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* .

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

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.