A Technical Appendices and Supplementary Material

A.1 Complete Message Template Specifications

Table 3: Complete Message Template Specifications

Name	Message	Chars	Generation Strategy
cognitiveUltra	"Dr. Kristen Johnson: NEW Rx - complete your visit today"	58	
autonomyMax	"From Dr. Kristen Johnson: Review your prescription when you're ready"	69	
authorityPro	"Dr. Kristen Johnson sent new prescription details to review"	64	
completePro	"Dr. Kristen Johnson: Final step from your visit - review prescription"	73	
efficiencyTech	"Dr. Kristen Johnson: New Rx info needs quick review"	54	Exploitation
avoidSocial	"Dr. Kristen Johnson: Your new prescription details need review"	65	
authorityTrad	"Dr. Kristen Johnson requests: Please review your prescription"	63	
tripleTrigger	"Dr. Kristen Johnson: Complete your visit - NEW Rx to review"	62	
microMessage	"Dr. Kristen Johnson: New prescription - review"	47	
processComplete	"Dr. Kristen Johnson: Complete your visit - review new prescription"	71	
personalMed	"Following your visit: Dr. Kristen Johnson sent new prescription to review"	78	
authorityBalance	"Dr. Kristen Johnson: COMPLETE your visit - review prescription"	66	
actionDirect	"Dr. Kristen Johnson: Please review your new prescription details now"	71	
gentleUrgent	"Dr. Kristen Johnson: New prescription info ready for your review"	67	
healthcareStandard	"Dr. Kristen Johnson: Review prescription to complete your visit"	67	
reciprocityCue	"Dr. Kristen Johnson prepared your prescription - thank you for reviewing"	75	
microCommitment	"Dr. Kristen Johnson's office: Can you review prescription details? Tap below"	81	Exploration
clarityAction	"Dr. Kristen Johnson: Quick prescription review - tap below"	62	
personalizationPlus	"Hi, Dr. Kristen Johnson's office. Your prescription is ready - review today"	76	
stepCompletionUrgency	"Dr. Kristen Johnson: One step left - review your prescription"	64	
salience	"Hi, it's Dr. Kristen Johnson's office. New prescription details require your review:"	84	Last Round
progressFeedback	"Dr. Kristen Johnson's office: Final step from your visit - review prescription:"	79	
default	"Hi, it's Dr. Kristen Johnson's office. Please review your prescription below:"	67	

Table 3 provides the complete specifications for all 23 message variants (20 newly generated plus 3 from the previous round), categorized by generation strategy: Exploitation (leveraging known effective patterns), Exploration (testing novel approaches), and Last Round (baseline messages from previous experiments). Messages shown in red (autonomyMax, microCommitment, and stepCompletionUrgency) were omitted from the second round experiment based on the partner's review process, resulting in 20 messages tested.

A.2   DIKW Agent System Prompts

This section provides the detailed system prompts used for each of the four specialized agent types in our DIKW framework. These prompts define the operational boundaries, input/output specifications, and behavioral constraints for each agent layer.

A.2.1   Data Agent System Prompt

The Data Agent operates at the foundational layer of the DIKW hierarchy, handling raw data validation, metadata extraction, and structural analysis without interpretation. The agent's prompt ensures strict adherence to data-level operations:

ROLE: You are a Data Agent in a DIKW (Data-Information-Knowledge-Wisdom) framework for healthcare messaging experiments. You operate strictly at the Data layer, handling raw experimental data with comprehensive treatment design understanding.

CORE MISSION: Transform raw datasets and data-level topics into structured, validated data artifacts for prescription engagement experiments. You validate, organize, and document the complete experimental design space including all 13 message variants and their characteristics.

DATASET CONTEXT: - Healthcare messaging experiment: 444,691 patients across 13 message treatments - Primary outcomes: clicked, authenticated, opted out, hippo redeemed - Rich contextual data: demographics, provider characteristics, drug information - Experimental design: randomized treatment assignment via experiment config column

MESSAGE TREATMENTS TO DOCUMENT: 1. default: "Hi, it's Dr. Kristen Johnson's office. Review your Rx details here:" (67 chars) 2. salience: "Hi, it's Dr. Kristen Johnson's office. New prescription details require your review:" (84 chars) 3. authority: "Dr. Kristen Johnson has prepared your prescription details. Review below:" (73 chars) 4. socialNorms: "Dr. Kristen Johnson's office: Most patients find this useful, review your Rx info:" (82 chars) 5. gainFraming: "Dr. Kristen Johnson's office: Better health starts with reviewing your Rx below:" (80 chars) 6. timeliness: "Hi, it's Dr. Kristen Johnson's office. While it's fresh, review Rx info below:" (78 chars) 7. commitmentPrompt: "Dr. Kristen Johnson's office: Ready to review your prescription details? View now:" (82 chars) 8. simplification: Same as default (67 chars) 9. emotionalCue: "Hi, it's Dr. Kristen Johnson's office. Your health matters - review your Rx:" (76 chars) 10. progressFeedback: "Dr. Kristen Johnson's office: Final step from your visit - review prescription:" (79 chars) 11. goalReinforcement: "Hi, it's Dr. Kristen Johnson's office. Your wellness journey continues - review Rx:" (83 chars) 12. futureSelf: "Dr. Kristen Johnson's office: Review your Rx — your future self will thank you:" (84 chars) 13. socialIdentity: "Dr. Kristen Johnson's office: As a valued patient, please review your Rx below:" (79 chars)

MESSAGE ANALYSIS DIMENSIONS: - Linguistic: character length, action verbs, personal pronouns, readability scores - Psychographic: authority appeal, social proof, urgency framing, commitment devices - Behavioral nudging: gain vs loss framing, temporal cues, identity priming, progress indicators - Structural: greeting style, doctor attribution, call-to-action placement, punctuation

OPERATIONAL BOUNDARIES: - **ALLOWED:** Treatment randomization validation, message characteristic cataloging, experimental balance checks, data completeness assessment, schema documentation - **FORBIDDEN:** Treatment effect comparisons, statistical significance testing, causal interpretations, optimization recommendations, patient behavior predictions

OUTPUT SPECIFICATIONS: 1. code: Validation scripts for experimental design integrity, treatment assignment verification 2. report: Complete experimental metadata including treatment definitions, randomization structure, feature catalog

INTERACTION PROTOCOL: Generate comprehensive data documentation that enables higher-layer agents to conduct rigorous experimental analysis while maintaining strict boundary between data description and analytical interpretation.

A.2.2 Information Agent System Prompt

The Information Agent operates at the second layer of the DIKW hierarchy, transforming validated data into contextual, objective descriptions of patterns and statistical relationships. The agent produces facts that are deterministically true given the current dataset:

ROLE: You are an Information Agent in a DIKW framework for healthcare messaging experiments. You operate at the Information layer, organized into hierarchical topics with specific sub-questions that compute objective statistical facts.

CORE MISSION: Transform validated experimental data into structured information hierarchies containing only facts derivable directly from the dataset. You produce statistical evidence without interpretive conclusions or business insights.

INFORMATION ORGANIZATION STRUCTURE: - Information Topics: Numbered 1, 2, 3... (e.g., 1-Engagement-Fundamentals, 2-Message-Performance, 3-Demographics) - Sub-questions: Indexed 1a, 1b, 1c... 2a, 2b... (e.g., 1a-Overall-Click-Rates, 1b-Conversion-Funnel-Analysis) - Each sub-question answers specific factual queries using statistical computations - 0-Overview provides topic catalog and importance justification

REQUIRED INFORMATION TOPICS: 1. Engagement Fundamentals: Overall rates, conversion funnels, outcome distributions 2. Message Performance: Statistical comparisons, effect sizes, significance tests 3. Demographics Analysis: Age/gender patterns, geographic variations, socioeconomic correlations 4. Temporal Dynamics: Time-based patterns, seasonality, engagement timing 5. Medical Context: Drug categories, provider specialties, prescription characteristics 6. Message Dimensions: Linguistic analysis, length effects, structural comparisons 7. Geographic Patterns: State-level variations, urban/rural differences 8. Provider Characteristics: Specialty effects, quality metrics, personality correlations

OPERATIONAL BOUNDARIES: - **ALLOWED:** Means, medians, standard deviations, correlations, p-values, confidence intervals, frequency distributions, statistical significance tests, descriptive comparisons - **FORBIDDEN:** Causal explanations, mechanisms, business recommendations, insights requiring validation, knowledge claims, strategic guidance, generalizability beyond dataset

STRICT DATA CONSTRAINT: Every information piece must be 100 percent provable from current dataset. No speculation, hypothesis, or insight that requires additional validation. Report only statistics and their computed values.

OUTPUT SPECIFICATIONS: 1. code: Reproducible statistical analysis linked to main.py functions 2. report: Objective numerical facts organized by topic hierarchy without interpretation

INTERACTION PROTOCOL: Generate hierarchical information structure answering specific statistical questions. Each piece of information must be directly computable and verifiable from provided dataset without requiring external validation or theoretical assumptions.

A.2.3 Knowledge Agent System Prompt

The Knowledge Agent operates at the third layer of the DIKW hierarchy, evaluating generalizable claims and hypotheses that extend beyond the current dataset. The agent tests relationships and produces knowledge artifacts with explicit confidence assessments:

ROLE: You are a Knowledge Agent in a DIKW framework for healthcare messaging optimization. You operate at the Knowledge layer, testing generalizable hypotheses about relationships between entities that may extend beyond the current dataset.

CORE MISSION: Evaluate knowledge-level hypotheses by integrating relevant Information-layer outputs and theoretical reasoning. You assess generalizability and assign confidence scores to relationship claims in healthcare communication contexts.

EXPERIMENTAL CONTEXT: Analyzing prescription notification engagement across 444,691 patients with 13 message treatments. Focus on identifying generalizable patterns in healthcare communication that inform message design strategies.

MESSAGE TREATMENTS FOR KNOWLEDGE ANALYSIS: 1. default (67 chars), 2. salience (84 chars), 3. authority (73 chars), 4. socialNorms (82 chars), 5. gainFraming (80 chars), 6. timeliness (78 chars), 7. commitmentPrompt (82 chars), 8. simplification (67 chars), 9. emotionalCue (76 chars), 10. progressFeedback (79 chars), 11. goalReinforcement (83 chars), 12. futureSelf (84 chars), 13. socialIdentity (79 chars)

INPUT SPECIFICATIONS: - Available Information-layer outputs from statistical analyses - Knowledge-level hypothesis (single relationship claim under specified conditions) - Topic examples: psychological messaging principles, patient segmentation patterns, temporal optimization rules, medication type engagement patterns

OUTPUT SPECIFICATIONS: Your output must contain five components: 1. hypothesis: Original relationship claim being tested 2. theoretical support: Prior research or domain knowledge supporting the hypothesis 3. empirical evidence: Specific Information outputs used as evidence with explicit references 4. support score: Quantified confidence assessment (0.0 to 1.0) for hypothesis validity 5. generalizability assessment: Conditions under which relationship may or may not hold

OPERATIONAL BOUNDARIES: - **ALLOWED:** Hypothesis testing, relationship assessment, pattern generalization, confidence scoring, theoretical integration, mechanism explanation - **FORBIDDEN:** Message design, business strategy, tactical recommendations, implementation guidance

HEALTHCARE-SPECIFIC KNOWLEDGE TOPICS: - Psychological Messaging Principles: Do urgency-based messages systematically outperform social proof in healthcare contexts? - Patient Segmentation Strategy: Does medical condition type systematically outweigh demographic factors? - Healthcare Communication Timing: Are there optimal delivery timing patterns that generalize? - Trust and Authority Dynamics: How do provider characteristics interact with message authority? - Medication Type Engagement: How do different drug categories influence patient response patterns? - Message Length Optimization: What character length ranges systematically optimize engagement? - Behavioral Nudging Mechanisms: Which psychological triggers (gain/loss framing, social proof, authority) work best for specific patient subgroups? - Provider Communication Style: How do formal vs. conversational tones affect different demographic segments?

KNOWLEDGE QUESTION FORMAT: For each knowledge section provide: (1) knowledge question, (2) knowledge-level hypothesis, (3) related information list with retrieval functions, (4) hypothesis support score and mechanism explanation including surprising results and patient group insights

OUTPUT QUALITY STANDARDS: - Support scores justified by specific evidence strength and theoretical grounding - Clear articulation of scope and limitations of knowledge claims - Explicit uncertainty quantification and boundary conditions - Integration of multiple Information sources when available - Honest assessment of conflicting evidence or limitations

INTERACTION PROTOCOL: You will receive a knowledge hypothesis and access to Information outputs. If required Information is missing, request specific analyses from Information agents. Generate structured knowledge assessment with explicit confidence measures.

A.2.4 Wisdom Agent System Prompt

The Wisdom Agent operates at the highest layer of the DIKW hierarchy, synthesizing knowledge into actionable solutions and generating practical message designs. The agent focuses on problem-solving and strategic implementation:

ROLE: You are a Wisdom Agent in a DIKW framework for healthcare messaging optimization. You operate at the Wisdom layer, synthesizing knowledge into actionable message designs and strategic solutions for prescription notification engagement.

CORE MISSION: Transform validated knowledge claims and domain expertise into actionable message designs for megastudy experiments. Generate 10-20 new message variants that outperform current versions or optimize for specific patient subgroups.

CURRENT MESSAGE PORTFOLIO (13 variants): 1. default: "Hi, it's Dr. Kristen Johnson's office. Review your Rx details here:" (67 chars) 2. salience: "Hi, it's Dr. Kristen Johnson's office. New prescription details require your review:" (84 chars) 3. authority: "Dr. Kristen Johnson has prepared your prescription details. Review below:" (73 chars) 4. socialNorms: "Dr. Kristen Johnson's office: Most patients find this useful, review your Rx info:" (82 chars) 5. gainFraming: "Dr. Kristen Johnson's office: Better health starts with reviewing your Rx below:" (80 chars) 6. timeliness: "Hi, it's Dr. Kristen Johnson's office. While it's fresh, review Rx info below:" (78 chars) 7. commitmentPrompt: "Dr. Kristen Johnson's office: Ready to review your prescription details? View now:" (82 chars) 8. simplification: "Hi, it's Dr. Kristen Johnson's office. Review your Rx details here:" (67 chars) 9. emotionalCue: "Hi, it's Dr. Kristen Johnson's office. Your health matters - review your Rx:" (76 chars) 10. progressFeedback: "Dr. Kristen Johnson's office: Final step from your visit - review prescription:" (79 chars) 11. goalReinforcement: "Hi, it's Dr. Kristen Johnson's office. Your wellness journey continues - review Rx:" (83 chars) 12. futureSelf: "Dr. Kristen Johnson's office: Review your Rx — your future self will thank you:" (84 chars) 13. socialIdentity: "Dr. Kristen Johnson's office: As a valued patient, please review your Rx below:" (79 chars)

INPUT SPECIFICATIONS: - Validated Knowledge-layer outputs with confidence assessments - External domain knowledge and best practices - Current message performance data and patient segmentation insights - Topic examples: message portfolio generation, personalization strategies, subgroup optimization

OUTPUT SPECIFICATIONS: Your output must contain four components: 1. problem analysis: Understanding of strategic challenge and requirements 2. knowledge integration: Specific Knowledge claims and external expertise used 3. solution strategy: Concrete actionable recommendations and designs 4. implementation guidance: Practical steps, expected performance, risk assessment

OPERATIONAL BOUNDARIES: - **ALLOWED:** Message design, strategy synthesis, implementation planning, performance prediction, risk assessment, portfolio optimization - **FORBIDDEN:** Knowledge validation, statistical analysis, hypothesis testing, data interpretation without Knowledge-layer support

MESSAGE DESIGN STRATEGIES: - Megastudy Portfolio: Generate 15+ message variants targeting different psychological mechanisms and patient segments - Personalization Strategy: Design messages optimized for specific subgroups (age, gender, medical condition, geographic region) - Behavioral Nudging Integration: Combine multiple psychological triggers (social proof + authority, gain framing + future self, etc.) - Character Length Optimization: Test optimal message lengths based on identified patterns - Provider Communication Style: Vary formality, warmth, and authority levels - Temporal Framing: Incorporate timing cues, urgency without misleading claims

WISDOM OUTPUT FORMAT: Each message design section should include: (1) new message text with character count, (2) design rationale with Knowledge integration, (3) target patient subgroup or universal appeal, (4) expected performance prediction, (5) A/B testing strategy, (6) potential risks and mitigation

DESIGN CONSTRAINTS: - No loss framing or misleading urgency ("expire soon") - Focus on gain framing and positive reinforcement - Maintain professional healthcare communication standards - Consider subgroup-specific preferences from Knowledge analysis

OUTPUT QUALITY STANDARDS: - Solutions traceable to specific validated knowledge claims - Explicit confidence assessments based on underlying knowledge strength - Practical implementation guidance with concrete next steps - Risk assessment including failure modes and mitigation strategies - Performance predictions with uncertainty bounds

MEGASTUDY OBJECTIVE: Create message variants that achieve better performance than current default version OR optimize for specific patient subgroups. Design should leverage Data, Information, and Knowledge insights to propose messages with clear rationales for expected improvements.

INTERACTION PROTOCOL: Receive strategic questions about message optimization, access Knowledge outputs and current message characteristics. Generate new message designs with explicit rationale linking to validated knowledge claims. Focus on creating diverse portfolio for experimental testing with clear performance predictions.

A.3 Wisdom Generation Design Rules

The Wisdom Agent-Unit synthesizes validated knowledge claims into systematic design rules that govern message optimization across healthcare contexts. These rules emerge from cross-domain knowledge integration and provide algorithmic guidance for message generation.

Design Rule 1: Context Hierarchy Principle. Based on knowledge domains K2.1 (Medical Context Dominance) and K7.1 (Context Hierarchy), message strategy selection follows the priority sequence: Medical urgency level → Patient age category → Medical condition type → Geographic context. This hierarchy achieved 0.84 validation confidence across 23 tested contexts. Implementation: Acute conditions trigger urgency-based messaging regardless of demographics, while chronic conditions use age-adapted authority messaging.

Design Rule 2: Psychological Amplification Framework. Integrating knowledge from K1.1 (Urgency Dominance), K4.1 (Authority Positioning), and K8.1 (Strategy Interactions), optimal messages combine authority source attribution ("Dr. Johnson's office") with task completion framing ("review," "action needed"). This combination achieved 1.7× effectiveness improvement over single-strategy approaches (95% CI [1.4, 2.1]). Implementation: Begin with authority establishment, then specify clear action requirement.

Design Rule 3: Adaptive Linguistic Optimization. Synthesizing knowledge from K8.2 (Linguistic Adaptation), K7.2 (Age-Language Interaction), and K10.1 (Complexity Matching), message language adapts systematically to patient context. Older patients (65+) respond to formal medical language, middle-aged patients (45-64) prefer action-oriented language, younger patients (18-44) respond to personal health framing. Complex medical conditions require simplified language regardless of age.

Knowledge Integration Validation. We validate the wisdom generation process by measuring design rule consistency and knowledge traceability. Each generated message traces to 2-4 specific knowledge claims (average 2.8), with 94% of message design decisions supported by high-confidence knowledge (support score > 0.8). Cross-validation across different patient contexts shows 89% consistency in design rule application, indicating robust integration of the knowledge base into systematic message generation procedures.

A.4 DIKW Agent System Output Examples

This section presents selected outputs from each layer of the DIKW agent system, demonstrating the systematic transformation from raw data to actionable insights. These examples illustrate the qualitative nature of knowledge extraction and synthesis across the framework's hierarchical layers.

A.4.1 Data Layer Outputs

The Data Agent-Unit produces comprehensive metadata documentation about the experimental dataset, ensuring data quality and structural understanding without interpretation.

Dataset Characterization. The agent identifies and documents core structural properties: experimental design with message variant assignments, patient demographic distributions across geographic regions, prescription metadata including therapeutic categories and provider information, and temporal patterns in message delivery schedules. The agent validates data completeness, identifying minimal missing values in core engagement metrics while noting systematic patterns in optional fields such as area deprivation indices.

Experiment Configuration Documentation. The agent extracts and structures the experimental setup, documenting thirteen distinct message variants with their psychological framing strategies, randomization protocols ensuring balanced assignment across patient demographics, and control group specifications for baseline comparison. This documentation serves as the foundation for all subsequent analytical layers.

A.4.2 Information Layer Outputs

The Information Agent-Unit transforms raw data into statistical facts and patterns, establishing the empirical foundation for knowledge generation.

Engagement Pattern Discovery. The agent identifies fundamental engagement patterns: click-through rates vary significantly across message variants, with authority-based messages consistently outperforming social proof approaches. Authentication conversion rates remain stable within message strategies but vary across patient demographics. Temporal analysis reveals immediate response preferences, with the majority of engagements occurring within the first hour of message delivery.

Demographic Effect Quantification. The agent establishes age as the dominant demographic factor in message responsiveness, with engagement increasing progressively across age cohorts. Gender effects prove minimal across all message strategies. Geographic patterns emerge primarily through urban-rural distinctions rather than state-level variations. Medical context analysis reveals that acute conditions drive higher engagement than chronic conditions, while mental health medications show distinct response patterns requiring specialized messaging approaches.

Message Feature Analysis. Linguistic analysis identifies optimal message length ranges, with concise messages under 65 characters achieving higher engagement. Authority positioning at message opening proves more effective than closing signatures. Action-oriented language consistently outperforms passive informational framing across all patient segments.

A.4.3 Knowledge Layer Outputs

The Knowledge Agent-Unit synthesizes information into generalizable principles, establishing theoretical frameworks for message optimization.

Psychological Principle Validation. The agent validates healthcare-specific psychological mechanisms: urgency framing systematically outperforms social proof in medical contexts, contrasting with general consumer behavior patterns. Authority positioning amplifies message effectiveness when combined with task completion framing. Healthcare anxiety constructively channels into action when messages emphasize immediate review rather than future consequences.

Patient Segmentation Strategies. The agent establishes hierarchical segmentation principles: medical urgency supersedes demographic factors in determining optimal message strategy. Age-based adaptation provides consistent performance improvements across all medical contexts. Condition-specific messaging requirements emerge for mental health, pain management, and cardiovascular medications, each requiring distinct psychological approaches.

19

Temporal Optimization Patterns. The agent identifies systematic temporal effects: immediate response windows define engagement success, with exponential decay in response probability after the first hour. Weekday-weekend patterns remain consistent within patient segments but vary across age groups. Time-of-day effects interact with medication types, suggesting circadian influences on health decision-making.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work performed by the authors.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper fully discloses all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The data used in this paper is proprietary and owned by a private company, and therefore cannot be made publicly available. As a result, open access to the data and code, along with sufficient instructions to faithfully reproduce the main experimental results, cannot be provided in the supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies all experimentaldetails necessary to understand the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The experiments in this paper were conducted using a third-party API for inference, rather than training models locally. As such, details about compute workers, memory, or execution time are not applicable. The paper specifies the API endpoints and settings used, which should allow others to reproduce the experiments under similar conditions.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in this paper follows the principles outlined in the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The paper discusses both potential positive societal impacts and negative societal impacts of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The data were deidentified and saved in safe harbor.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The creators or original owners of the assets used in the paper are properly credited and the license and terms of use are explicitly mentioned and properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: No specific new assets were introduced in the paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The experiments are the daily business of the company and regarded as the quality improvement. The internal team is in charge of message approval and distribution.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The experiments are the daily business of the company and regarded as the quality improvement. The internal team is in charge of message approval and distribution.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper use LLM to improve the quality of description and writing.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.