

**NIST Technical Note
NIST TN 2361**

Validating LLM-Generated Data Grounded in Technical Documents

An Application in Community Planning

Juan F. Fung
Daniel K. Stephens
Alden Dima
Michael Majurski

This publication is available free of charge from:
<https://doi.org/10.6028/NIST.TN.2361>

**NIST Technical Note
NIST TN 2361**

**Validating LLM-Generated Data
Grounded in Technical Documents**

An Application in Community Planning

Juan F. Fung¹, Daniel K. Stephens², Alden Dima³, Michael Majurski³

¹ Applied Economics Office, Engineering Laboratory, NIST, Gaithersburg, Maryland, USA

² Department of Electrical Engineering, Morgan State University, Baltimore, Maryland, USA

³ Software and Systems Division, Information Technology Laboratory, NIST, Gaithersburg, Maryland, USA

This publication is available free of charge from:
<https://doi.org/10.6028/NIST.TN.2361>

February 2026



U.S. Department of Commerce
Howard Lutnick, Secretary

National Institute of Standards and Technology

Craig Burkhardt, Acting Under Secretary of Commerce for Standards and Technology and Acting NIST Director

NIST Technical Series Publications: <https://www.nist.gov/nist-research-library/nist-publications>

Non-Endorsement Disclaimer

Certain equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Generative AI Use

This manuscript was edited with the assistance of a locally-hosted instance of Llama-4-Maverick-17B-128E-Instruct-FP8, developed by Meta. Llama-4 was used to refine language, improve clarity, and enhance readability in accordance with the authors' instructions. All content, scientific claims, and conclusions have been reviewed and verified by the authors to ensure accuracy and originality.

Publication History

Approved by the NIST Editorial Review Board on 2026-01-21.

Suggested Citation

Juan F. Fung, Daniel K. Stephens, Alden Dima, Michael Majurski (2026). Validating LLM-Generated Data Grounded in Technical Documents: An Application in Community Planning. (National Institute of Standards and Technology, Gaithersburg, MD) NIST Technical Note (TN) NIST TN 2361.
<https://doi.org/10.6028/NIST.TN.2361>

Author Names and ORCID Identifiers

Juan F. Fung (0000-0002-0820-787X); Daniel K. Stephens (0009-0007-3122-7004); Alden Dima (0000-0003-0547-3117); Michael Majurski (0000-0001-9663-3803).

Contact Information: 2361-comments@list.nist.gov

Abstract

This report presents a novel approach to generating synthetic variations of technical community planning documents using Large Language Models (LLMs). We developed a comprehensive framework for transforming domain-specific resilience, adaptation, and sustainability (RAS) planning documents into more user-friendly versions while preserving semantic content. Our methodology employs multiple modification strategies including linguistic simplification, jargon removal, content augmentation, and tone adjustment, implemented through parallel processing pipelines. The system generates both individual and cumulative document modifications, which are then validated using a comprehensive framework that includes multiple similarity metrics. This validation framework provides a robust assessment of the semantic preservation achieved by the user-friendly modified documents. The process provides a template for generating and validating rich data while balancing user needs and maintaining technical accuracy in specialized domains.

Keywords

synthetic data generation, evaluation, validation, large language models (LLMs), community resilience planning.

Table of Contents

Abstract	i
List of Figures	iv
List of Tables	iv
Acknowledgments	iv
1. Introduction	1
1.1. Background and Literature	1
2. Generation Methodology	2
2.1. Data Source and Characteristics	2
2.2. Modification Strategies	2
2.3. Technical Implementation	3
2.4. Prompt Engineering Framework	4
2.5. System Performance and Scalability	4
3. Validation Methodology	5
3.1. Multi-Metric Semantic Similarity Framework	5
3.2. Statistical Analysis Framework	6
4. Results and Analysis	6
4.1. Overall Performance Summary	6
4.2. Similarity Scores Across Modification Strategies	6
4.3. Statistical Validation	7
4.4. Effect Sizes	9
4.5. Analysis of Modification Strategies	11
5. Limitations	12
6. Conclusions	13
References	14
Appendix A. Acronyms	16
Appendix B. Pseudo-Code for Synthetic Data	16
B.1. Task Definition	16
B.2. Document Processing	17
B.3. Corpus Generation	17

Generation

Appendix C. Document Examples	17
C.1. High Similarity Score	17
C.1.1. Original	17
C.1.2. Content Augmentation Modification (GSE Score: 0.995)	18
C.2. Low Similarity Score (Hallucination)	20
C.2.1. Original	20
C.2.2. Cumulative Modification (CTA: 0.377)	20
C.3. Failure to Generate a Response	21
C.3.1. Original	21
C.3.2. Jargon Modification (SSE Score: 0.073)	21

List of Figures

Figure 1. Similarity scores across modification strategies and metrics. Top row: Box plots showing the distribution of similarity scores for each modification strategy across General Semantic Embedding (GSE), Contextual Text Alignment (CTA), and Sentence-Level Semantic Encoder (SSE) metrics. Bottom row: Histograms showing the overall distribution of similarity scores for each metric.	7
Figure 2. Effect sizes (Cohen’s d) by modification and metric. GSE: General Semantic Embedding; CTA: Contextual Text Alignment; and SSE: Sentence-Level Semantic Encoder.	10
Figure 3. Similarity validation success rates by metric and modification type. Thresholds: GSE = 0.8; SSE = 0.7; and CTA = 0.6.	11

List of Tables

Table 1. Mean Similarity Scores and P-Values Across Modification Strategies and Metrics	8
Table 2. Summary of T-Test Results Across Modification Strategies and Metrics	8
Table 3. Summary of ANOVA Test Results Across Modification Strategies	9

Acknowledgments

We are grateful to David Butry, Christopher Clavin, Donghwan Gu, and Jason Averill for comments. All errors are our own.

1. Introduction

Traditional approaches to document simplification rely on rule-based systems or require extensive manual effort. This work explores the potential of Large Language Models (LLMs) to automatically generate accessible versions of technical planning documents while maintaining substantive content integrity.¹

The research emerges from practical challenges encountered in content analysis of community planning documents, where domain-specific terminology and formal language create barriers to broader stakeholder engagement. Community resilience planning documents in particular are critical resources for disaster preparedness, yet their technical nature often limits accessibility to non-expert stakeholders. Our approach transforms this limitation into an opportunity by systematically generating synthetic document variants that maintain core meaning while improving accessibility, validated through state-of-the-art semantic similarity metrics.

1.1. Background and Literature

The use of LLMs to generate synthetic data is growing, with applications including text data for training and fine-tuning language models, instructions for language model inference, and synthetic images for training computer vision [1, 2]. Often the goal is to improve model performance on a task [3–5]. A key challenge is ensuring synthetic data is of high quality and relevant for the task at hand. Methods for improving the quality of generated data include sampling from multiple prompts and generating desired document attributes that can be parameterized (e.g., topic, writing style) [6]. But how do you measure the quality of the resulting synthetic documents?

Despite advances in synthetic data generation, there is no standard to evaluating synthetic data [1]. Many of the approaches can be traced to the literature on evaluating the quality of abstractive summarization [7–9]. Evaluations of synthetic data quality include factuality, fidelity, relevance to the task, and performance on benchmarks [6, 10–12]. Text entailment checks, for instance, are an approach to verifying the factual consistency of synthetic data with the ground truth data it is based on by assessing whether a statement is supported by a document [13–15]. Lupidi et al. [3] use the WikiSQL dataset to evaluate the performance of LLMs fine-tuned on synthetic data on tabular question-answering tasks. Other approaches include red teaming, human evaluation, and contamination detection [12]. Human evaluation may be the gold standard but is costly and does not scale [4, 7, 16, 17].

¹The code used in for synthetic data generation and validation in this report is available at: <https://github.com/juanfung/grounded-synthetic-data.git>

The common themes in the literature on generation and evaluation are that: the typical audience for synthetic data is a machine rather than a human; and evaluation is measured with respect to machine performance on a task. In contrast, the intended audience for the synthetic data in this report is a human, in particular one without domain expertise. Given the lack of a standardized approach to evaluating synthetic data in general, this report presents a first step toward building a comprehensive evaluation framework.

2. Generation Methodology

2.1. Data Source and Characteristics

The source dataset consists of community planning documents focused on resilience, adaptation, and sustainability (RAS) initiatives. These documents were previously annotated through manual content analysis and evaluated in user studies examining AI-assisted content analysis tools including active learning, topic modeling, and LLM-assisted labeling [18, 19]. The documents are characterized by:

- Domain-specific technical terminology
- Formal governmental/institutional language
- Complex policy and procedural content
- Variable document length and structure

The corpus comprises 1811 document excerpts characterized by domain specificity, institutional language, structural complexity, and annotation richness. The documents exhibit technical terminology related to community planning broadly and natural hazard risk reduction in particular, as well as formal governmental discourse. The structural complexity arises from variable document length (mean character length = 981, with a standard deviation of 1034), nested policy frameworks, and procedural content. The annotation richness is due to human-validated content labels enabling downstream validation.

2.2. Modification Strategies

We implemented ten distinct document modification approaches, beginning with three distinct **Linguistic Modifications**:

1. **Voice of America (VOA) English**: Simplification using broadcast journalism standards
2. **Jargon Removal**: Replacement of technical terms with accessible alternatives

3. **Content Augmentation:** Addition of explanatory text and definitions

We then considered the following **Stylistic Modifications:**

1. **Engagement Enhancement:** Improving reader engagement through stylistic changes
2. **Tone Modification:** Systematic modification to positive, negative, or neutral tones
3. **Topic Emphasis:** Amplification of primary document themes
4. **Theme Amplification:** Enhancement of key themes while minimizing secondary content

For Tone Modification, we applied each separately (positive, negative, neutral) for a total of six stylistic modifications. Finally, we considered a modification that combined each of the individual modifications, for a grand total of ten modifications:

Cumulative Modification: Sequential application of individual modification strategies, with neutral tone for the tone modification.

2.3. Technical Implementation

The modification strategies were implemented using asynchronous processing with concurrent execution, batch optimization, error resilience, rate management, and graceful degradation. The processing pipeline includes document ingestion, parallel modification, sequential processing, quality assurance, and structured export.

The synthetic data generation pipeline is idealized as follows:

1. **Individual Modifications:** Parallel execution of all modification strategies
2. **Cumulative Processing:** Sequential application of selected strategies
3. **Validation:** Automated consistency checking
4. **Output Generation:** Structured CSV export with all variants

The parallel processing system featured configurable batch sizes, exponential backoff retry logic, intelligent delays for API compliance, and continued processing despite individual modification failures. The system implements asynchronous processing with the following features:

- **Batch Processing:** Configurable batch sizes (default: 5-10 documents)
- **Retry Logic:** Exponential backoff with 3-attempt limit

- **Rate Limiting:** 5-second delays between batches
- **Error Handling:** Graceful degradation for API failures

Finally, we use a locally-deployed instance of an open-source LLM, with Temperature=0.7 and top_p=0.95.

2.4. Prompt Engineering Framework

All modification prompts followed a standardized template ensuring consistency. The core prompt architecture was tailored to make the document easy to understand for a layperson, with specific modification instructions for each strategy.

```
Your goal is to make this document {doc} easy to understand for  
a layperson.
```

```
Revise the document [specific modification instruction].
```

```
Return only the revised document.
```

In addition, we explored the use of a LLM-based validation. The validation prompt design employed comparative analysis prompts to assess consistency between original and modified documents. The prompts were formatted to provide a clear yes/no answer and explanation for the rationale.

Document consistency validation employs comparative analysis:

```
Compare a document without modifications, {doc_0}, and with  
modifications, {doc_1}, and validate that the modified document is 239  
consistent with the original document in terms of substance...
```

```
Format your response as:
```

- (Yes/No): Modified document is consistent with original
- Explanation: Explain your rationale

However, this was not a systematic approach to validation so we implemented an approach based on similarity scores that we will describe in Sec. 3.

2.5. System Performance and Scalability

The system demonstrated exceptional processing efficiency, with a throughput of 18 110 total transformations (1811 documents × 10 modifications) and a processing time of approximately 3.25 hours for the complete corpus transformation, with 100% success rate for all queries with an average of 0.65 seconds per query.

3. Validation Methodology

3.1. Multi-Metric Semantic Similarity Framework

To ensure rigorous evaluation of synthetic document quality, we implemented a validation framework employing multiple complementary similarity metrics. This approach addresses the fundamental challenge in synthetic data generation: quantifying whether automated modifications preserve the core semantic content while achieving usability improvements.

Our validation framework incorporates three similarity measurement approaches, each capturing different aspects of semantic preservation. This multi-metric approach triangulates similarity to minimize the bias of any single algorithm:

- General Semantic Embedding (GSE): A large-scale, general-purpose model that captures broad semantic equivalence and topical consistency (Metric: Cosine Similarity, Range: 0–1).
- Contextual Token Alignment (CTA): A precision-oriented algorithm that utilizes token-level matching to verify the retention of specific terms and local context (Metric: F1 Score, Range: 0–1).
- Sentence-Level Semantic Encoder (SSE): A specialized model fine-tuned to optimize sentence-level clustering and structural similarity (Metric: Cosine Similarity, Range: 0–1).

We use generic acronyms for the algorithms since the purpose of the study is not to compare any particular algorithm but rather to illustrate an approach to validation.

We establish metric-specific thresholds to evaluate the performance of our modification strategies. The thresholds were set as follows:

- $GSE \geq 0.8$
- $SSE \geq 0.7$
- $CTA \geq 0.6$

These thresholds reflect our expectations for strong semantic preservation using high-quality semantic embeddings, good semantic understanding at the sentence level, and lower contextual similarity, respectively. We then evaluate which modifications exceed the expected thresholds on average across documents.

3.2. Statistical Analysis Framework

To ensure statistical rigor, we conduct one-sided one-sample t -tests to test whether the mean similarity scores across documents per modification exceed the metric-specific thresholds. This analysis allows us to determine whether the observed similarity scores were statistically significant.

We also conduct ANOVA tests to compare similarity scores across metrics for a given modification. This analysis helps us understand whether the results are consistent across different metrics.

Finally, to assess the practical significance of our findings, we calculate effect sizes using Cohen's d . This analysis provides insight into the magnitude of the differences between the observed similarity scores and the metric-specific thresholds.

By using a combination of statistical tests and effect size calculations, we are able to provide a comprehensive evaluation of our modification strategies and demonstrate the effectiveness of synthetic data generation in preserving the substance from the original documents.

4. Results and Analysis

Our comprehensive validation analysis demonstrates that the modification strategies achieve high semantic preservation across all metrics, with nearly all modification strategies significantly exceeding the metric-specific thresholds. This section presents the results of the analysis, including the similarity scores across modification strategies, statistical validation, and modification strategy performance analysis.

4.1. Overall Performance Summary

The comprehensive validation analysis demonstrates that the modification strategies achieve high semantic preservation across all metrics. The results of the one-sided one-sample t -tests show that 9 out of 10 modification strategies significantly exceed the metric-specific thresholds, with a mean effect size of 2.66.

4.2. Similarity Scores Across Modification Strategies

The box plots in Fig. 1 provide a visual representation of the similarity scores across different modification strategies and metrics. The results show that most modification strategies

achieve high similarity scores across all metrics, with some variation in performance depending on the specific strategy and metric used.

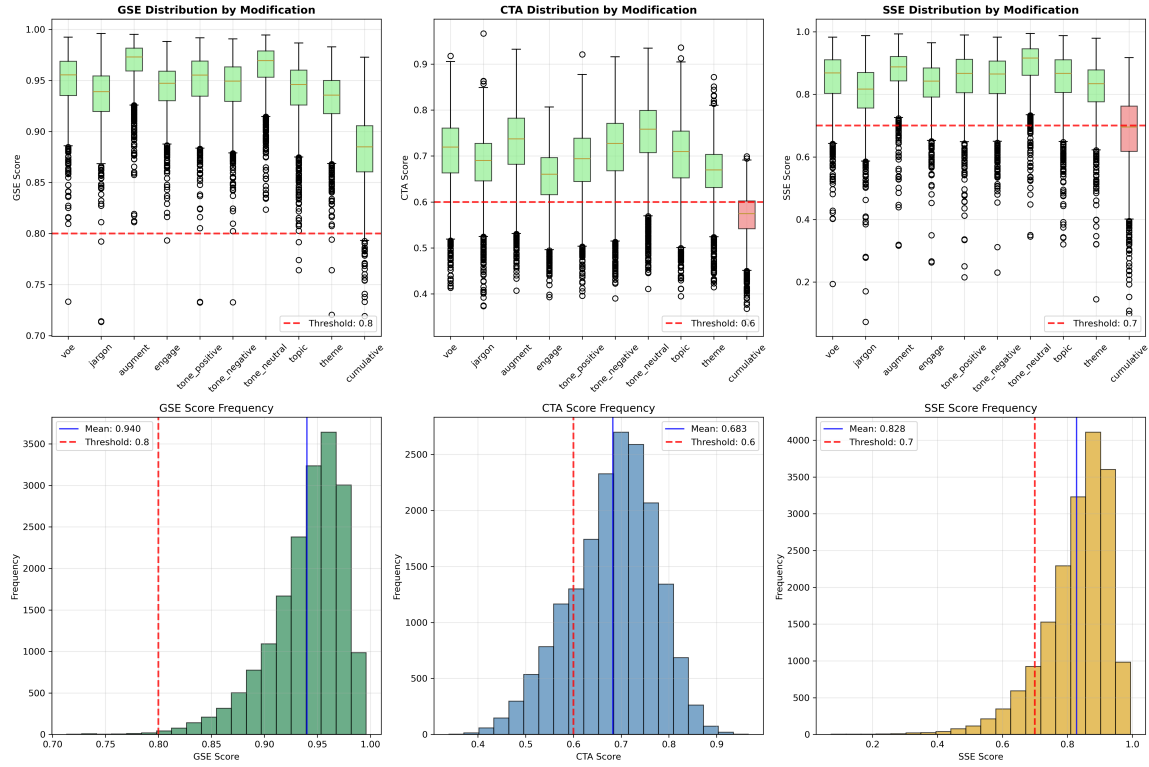


Figure 1. Similarity scores across modification strategies and metrics. Top row: Box plots showing the distribution of similarity scores for each modification strategy across General Semantic Embedding (GSE), Contextual Text Alignment (CTA), and Sentence-Level Semantic Encoder (SSE) metrics. Bottom row: Histograms showing the overall distribution of similarity scores for each metric.

Table 1 summarizes mean scores and p-values across metrics and modifications.

4.3. Statistical Validation

Building on the observed similarity scores, our statistical analysis provides compelling evidence that the LLM-based document modification approach can achieve both usability improvements and semantic fidelity. The results of the one-sided one-sample t-tests and the ANOVA tests demonstrate that the modification strategies achieve statistically significant similarity scores across all metrics. The effect size calculations using Cohen’s d provide insight into the magnitude of the differences between the observed similarity scores and the metric-specific thresholds.

Table 1. Mean Similarity Scores and P-Values Across Modification Strategies and Metrics

Modification Strategy	GSE	CTA	SSE
VOA Simplification	0.95 (p < 0.01)	0.71 (p < 0.01)	0.85 (p < 0.01)
Jargon Removal	0.94 (p < 0.01)	0.68 (p < 0.01)	0.80 (p < 0.01)
Content Augmentation	0.97 (p < 0.01)	0.72 (p < 0.01)	0.87 (p < 0.01)
Engagement Enhancement	0.94 (p < 0.01)	0.65 (p < 0.01)	0.83 (p < 0.01)
Positive Tone	0.95 (p < 0.01)	0.69 (p < 0.01)	0.85 (p < 0.01)
Negative Tone	0.94 (p < 0.01)	0.71 (p < 0.01)	0.84 (p < 0.01)
Neutral Tone	0.96 (p < 0.01)	0.74 (p < 0.01)	0.89 (p < 0.01)
Topic Emphasis	0.94 (p < 0.01)	0.70 (p < 0.01)	0.85 (p < 0.01)
Theme Amplification	0.93 (p < 0.01)	0.66 (p < 0.01)	0.82 (p < 0.01)
Cumulative Modifications	0.88 (p < 0.01)	0.57 (p=1.00)	0.68 (p=1.00)

The results of the one-sided one-sample t-tests in Table 2 show that 9 out of 10 modification strategies significantly exceed the metric-specific thresholds. The exception is the cumulative modification strategy, for which the CTA and SSE tests fail to reject the null.

Table 2. Summary of T-Test Results Across Modification Strategies and Metrics

Modification Strategy	GSE	CTA	SSE
VOA Simplification	p < 0.01	p < 0.01	p < 0.01
Jargon Removal	p < 0.01	p < 0.01	p < 0.01
Content Augmentation	p < 0.01	p < 0.01	p < 0.01
Engagement Enhancement	p < 0.01	p < 0.01	p < 0.01
Positive Tone	p < 0.01	p < 0.01	p < 0.01
Negative Tone	p < 0.01	p < 0.01	p < 0.01
Neutral Tone	p < 0.01	p < 0.01	p < 0.01
Topic Emphasis	p < 0.01	p < 0.01	p < 0.01
Theme Amplification	p < 0.01	p < 0.01	p < 0.01
Cumulative Modifications	p < 0.01	p=1.00	p=1.00

The ANOVA tests, summarized in Table 3, reveal significant differences between the metrics for all 10 modification strategies, indicating that the results are not consistent across different metrics. This finding suggests that the choice of metric has a significant impact on the evaluation of the modification strategies. For example, the mean similarity scores for GSE are generally higher than those for CTA, indicating that GSE may be more lenient in its evaluation of semantic similarity. In contrast, CTA appears to be more conservative, resulting in lower mean similarity scores.

Table 3. Summary of ANOVA Test Results Across Modification Strategies

Modification Strategy	F-Statistic	P-Value
VOA Simplification	5387.878	0.00
Jargon Removal	6037.223	0.00
Content Augmentation	6558.436	0.00
Engagement Enhancement	10845.452	0.00
Positive Tone	6860.600	0.00
Negative Tone	4709.758	0.00
Neutral Tone	5132.513	0.00
Topic Emphasis	5116.229	0.00
Theme Amplification	8154.853	0.00
Cumulative Modifications	7949.440	0.00

The significant differences between metrics highlight the importance of using multiple evaluation metrics to get a comprehensive understanding of the performance of the modification strategies. By considering multiple metrics, we can gain a more nuanced understanding of the strengths and weaknesses of each strategy and make more informed decisions about their application. The metrics we selected are complementary and capture different dimensions of semantic similarity.

The ANOVA results also have implications for the interpretation of the results. For instance, the significant differences between metrics suggest that the results should be interpreted in the context of the specific metric used. This means that the results may not be generalizable across different metrics, and caution should be exercised when comparing results across different studies or applications.

4.4. Effect Sizes

Finally, effect size calculations using Cohen's d provide insight into the magnitude of the differences between the observed similarity scores and the metric-specific thresholds. A heatmap of the effect sizes is shown in Fig. 2.

The results show that most modification strategies achieve large effect sizes across all metrics, indicating a significant difference between the observed similarity scores and the metric-specific thresholds. For example, Content Augmentation achieves an effect size of 7.61 for GSE, indicating a very large effect. Similarly, Neutral Tone achieves an effect size of 6.99 for GSE, also indicating a very large effect.

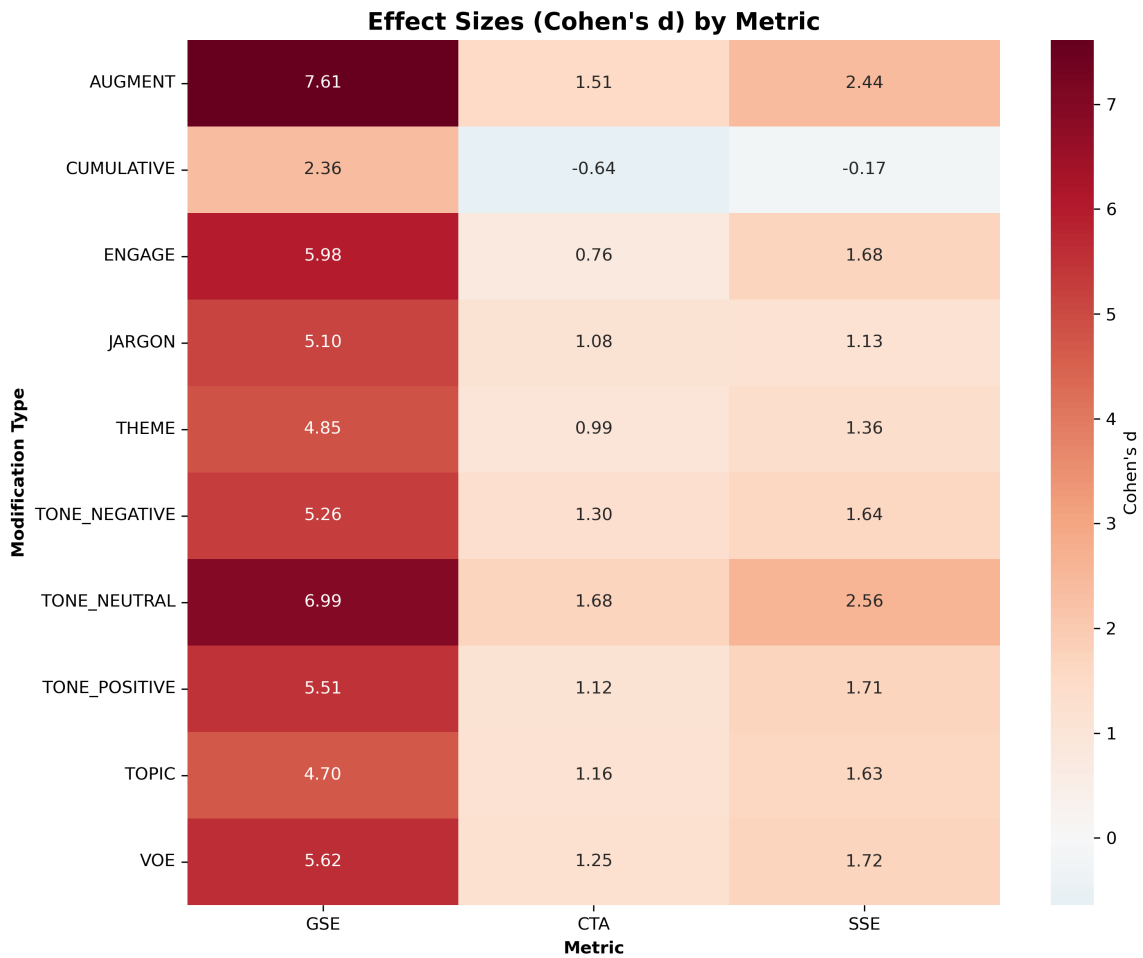


Figure 2. Effect sizes (Cohen's d) by modification and metric. GSE: General Semantic Embedding; CTA: Contextual Text Alignment; and SSE: Sentence-Level Semantic Encoder.

The effect sizes vary across metrics, with GSE generally showing larger effect sizes than CTA and SSE. This is consistent with the ANOVA results, which showed significant differences between the metrics.

The effect size calculations provide valuable insights into the practical significance of the results. While the statistical analysis provides evidence that the modification strategies achieve statistically significant similarity scores, the effect sizes indicate the magnitude of the differences between the observed scores and the thresholds. This information can be used to inform decisions about the application of the modification strategies and the potential benefits and limitations of the approach.

4.5. Analysis of Modification Strategies

A closer examination of the modification strategy performance reveals that Content Augmentation and Neutral Tone consistently achieve the highest similarity scores across all metrics. In contrast, Cumulative Modifications shows lower similarity scores, particularly for CTA. The heatmap in Fig. 3 provides a visual representation of the similarity validation success rates across different modification strategies and metrics, highlighting the strengths and weaknesses of each strategy. Appendix C illustrates examples of high and low similarity score modifications.

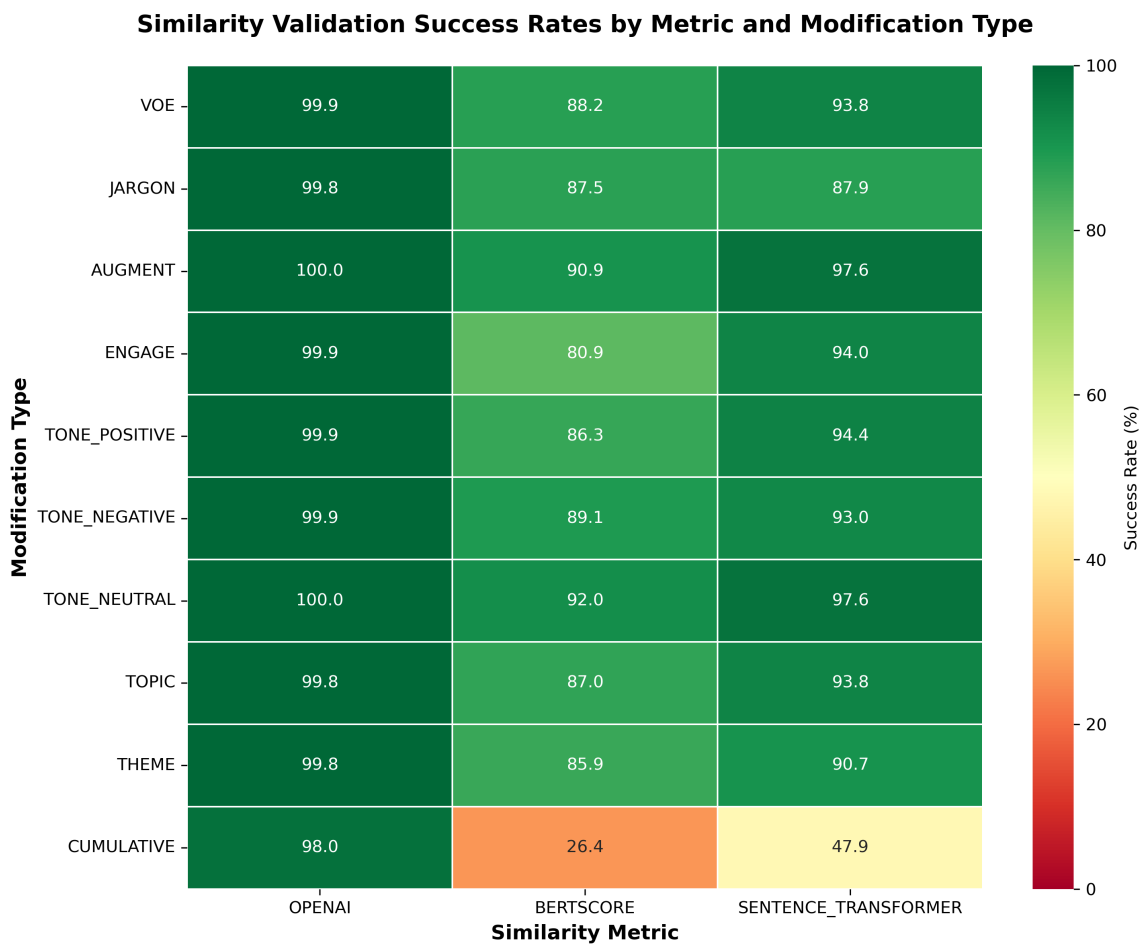


Figure 3. Similarity validation success rates by metric and modification type. Thresholds: GSE = 0.8; SSE = 0.7; and CTA = 0.6.

The performance of the different modification strategies varies across the three metrics. Content Augmentation and Neutral Tone Modification consistently achieve the highest similarity scores across all metrics, while Cumulative Modifications shows lower similarity

scores. Interestingly, low contextual (CTA) and sentence similarity scores revealed failed LLM responses for very short (about one sentence) documents, for which the LLM failed to generate a response (i.e., did not interpret the context provided as a document) or hallucinated a document with much more context than the provided document. See Appendix C for examples. The former were outliers and occurred for two documents, while the latter occurred for a few dozen short documents. In addition to providing a proxy for indirectly validating the LLM response succeeded, it highlights the importance of providing ground truth documents with sufficient context for modification.

Overall, the results demonstrate that the modification strategies achieve high semantic preservation across all metrics, with some variation in performance depending on the specific strategy and metric used. The statistical analysis provides a comprehensive understanding of the performance of the modification strategies and the potential applications of the approach.

5. Limitations

While the proposed approach demonstrates significant potential for generating synthetic variations of technical community planning documents, several limitations and considerations must be acknowledged. One key challenge is maintaining content fidelity, particularly in heavily modified documents where semantic drift may occur. As the degree of modification increases, there is a risk that the original meaning and intent of the document may be compromised.

Another limitation is the domain specificity of the approach, which is currently tailored to planning documents. While the methodology shows promise for this specific application, its generalizability to other domains or document types remains uncertain. Further research would be required to determine whether the approach can be effectively adapted for use in other contexts.

The current validation framework also has limitations, as it focuses primarily on consistency rather than comprehension. While the results demonstrate that the modified documents maintain a high level of semantic similarity to the originals, it is unclear whether readers can effectively comprehend the modified content. Future studies should investigate the comprehension and usability of the synthetic documents to provide a more comprehensive understanding of their value.

Finally, the resource requirements for large-scale processing are significant, and the computational costs associated with generating and validating synthetic documents should not

be underestimated. As the demand for synthetic data generation continues to grow, it will be essential to develop more efficient and scalable solutions to support this need.

6. Conclusions

This report presents a comprehensive framework for generating synthetic document variants using LLMs, specifically addressing the challenge of making technical planning documents more accessible. The methodology demonstrates the potential for LLM-based approaches to address real-world challenges in government communication and public engagement, while highlighting the need for rigorous validation frameworks to ensure content integrity across modification strategies.

The key contributions include the development of a comprehensive multi-metric validation framework, empirical validation of the approach through analysis of 1811 documents, and the establishment of a practical framework for quality assessment. The results demonstrate exceptional semantic preservation across all modification strategies, with mean similarity scores exceeding 88% for GSE.

The cross-metric convergence between three independent similarity metrics provides robust validation of the quality assessment approach and confirms the reliability of the synthetic data generation process. Clear performance hierarchies emerged across modification strategies, providing evidence-based guidance for future applications.

While the approach demonstrates success with our planning documents, generalization to other technical domains requires validation. Future research should examine performance across other technical documents in different domains to establish broader applicability. The cumulative modification results suggest that sequential transformations may compound semantic changes, and future work should investigate optimal transformation sequences and develop methods to minimize compound drift. Moreover, this represents a first step toward a more comprehensive evaluation framework that validates both semantic consistency with ground truth documents and quality improvements when the target reader of synthetic documents is a human. Finally, while the grounded synthetic data generation was one-to-one, future work should explore whether this approach is suited to one-to-many generation, also known as data augmentation, where the goal is to obtain a large data set from a small set of ground truth examples.

References

- [1] Bauer A, Trapp S, Stenger M, Leppich R, Kounev S, Leznik M, Chard K, Foster I (2024) Comprehensive exploration of synthetic data generation: A survey. 2401.02524 Available at <https://arxiv.org/abs/2401.02524>.
- [2] Guo X, Chen Y (2024) Generative ai for synthetic data generation: Methods, challenges and the future. 2403.04190 Available at <https://arxiv.org/abs/2403.04190>.
- [3] Lupidi A, Gemmell C, Cancedda N, Dwivedi-Yu J, Weston J, Foerster J, Raileanu R, Lomeli M (2025) Source2synth: Synthetic data generation and curation grounded in real data sources. 2409.08239 Available at <https://arxiv.org/abs/2409.08239>.
- [4] Wang Y, Kordi Y, Mishra S, Liu A, Smith NA, Khashabi D, Hajishirzi H (2023) Self-instruct: Aligning language models with self-generated instructions. 2212.10560 Available at <https://arxiv.org/abs/2212.10560>.
- [5] Zelikman E, Wu Y, Mu J, Goodman ND (2022) Star: Bootstrapping reasoning with reasoning. 2203.14465 Available at <https://arxiv.org/abs/2203.14465>.
- [6] Long L, Wang R, Xiao R, Zhao J, Ding X, Chen G, Wang H (2024) On llms-driven synthetic data generation, curation, and evaluation: A survey. 2406.15126 Available at <https://arxiv.org/abs/2406.15126>.
- [7] Kryscinski W, McCann B, Xiong C, Socher R (2020) Evaluating the factual consistency of abstractive text summarization. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, eds Webber B, Cohn T, He Y, Liu Y (Association for Computational Linguistics, Online), pp 9332–9346. DOI:10.18653/v1/2020.emnlp-main.750. Available at <https://aclanthology.org/2020.emnlp-main.750/>
- [8] Wang A, Cho K, Lewis M (2020) Asking and answering questions to evaluate the factual consistency of summaries. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, eds Jurafsky D, Chai J, Schluter N, Tetreault J (Association for Computational Linguistics, Online), pp 5008–5020. DOI:10.18653/v1/2020.acl-main.450. Available at <https://aclanthology.org/2020.acl-main.450/>
- [9] Tian K, Mitchell E, Yao H, Manning CD, Finn C (2023) Fine-tuning language models for factuality. 2311.08401 Available at <https://arxiv.org/abs/2311.08401>.
- [10] Min S, Krishna K, Lyu X, Lewis M, Yih Wt, Koh P, Iyyer M, Zettlemoyer L, Hajishirzi H (2023) FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, eds Bouamor H, Pino J, Bali K (Association for Computational Linguistics, Singapore), pp 12076–12100. DOI:10.18653/v1/2023.emnlp-main.741. Available at <https://aclanthology.org/2023.emnlp-main.741/>

- [11] Gupta H, Scaria K, Anantheswaran U, Verma S, Parmar M, Sawant SA, Baral C, Mishra S (2024) Targen: Targeted data generation with large language models. 2310.17876 Available at <https://arxiv.org/abs/2310.17876>.
- [12] Liu R, Wei J, Liu F, Si C, Zhang Y, Rao J, Zheng S, Peng D, Yang D, Zhou D, Dai AM (2024) Best practices and lessons learned on synthetic data. 2404.07503 Available at <https://arxiv.org/abs/2404.07503>.
- [13] Maynez J, Narayan S, Bohnet B, McDonald R (2020) On faithfulness and factuality in abstractive summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, eds Jurafsky D, Chai J, Schluter N, Tetreault J (Association for Computational Linguistics, Online), pp 1906–1919. DOI:10.18653/v1/2020.acl-main.173. Available at <https://aclanthology.org/2020.acl-main.173/>
- [14] Maini P, Seto S, Bai H, Grangier D, Zhang Y, Jaitly N (2024) Rephrasing the web: A recipe for compute and data-efficient language modeling. 2401.16380 Available at <https://arxiv.org/abs/2401.16380>.
- [15] Tang L, Laban P, Durrett G (2024) Minicheck: Efficient fact-checking of llms on grounding documents. 2404.10774 Available at <https://arxiv.org/abs/2404.10774>.
- [16] Du X, Shao J, Cardie C (2017) Learning to ask: Neural question generation for reading comprehension. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds Barzilay R, Kan MY (Association for Computational Linguistics, Vancouver, Canada), pp 1342–1352. DOI:10.18653/v1/P17-1123. Available at <https://aclanthology.org/P17-1123/>
- [17] Xu C, Sun Q, Zheng K, Geng X, Zhao P, Feng J, Tao C, Lin Q, Jiang D (2025) Wizardlm: Empowering large pre-trained language models to follow complex instructions. 2304.12244 Available at <https://arxiv.org/abs/2304.12244>.
- [18] Li Z, Mao A, Stephens D, Goel P, Walpole E, Dima A, Fung J, Boyd-Graber J (2024) Improving the TENOR of labeling: Re-evaluating topic models for content analysis. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds Graham Y, Purver M (Association for Computational Linguistics, St. Julian’s, Malta), pp 840–859. DOI:10.18653/v1/2024.eacl-long.51. Available at <https://aclanthology.org/2024.eacl-long.51/>
- [19] Li Z, Calvo-Bartolomé L, Hoyle AM, Xu P, Stephens DK, Fung JF, Dima A, Boyd-Graber JL (2025) Large language models struggle to describe the haystack without human help: A social science-inspired evaluation of topic models. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds Che W, Nabende J, Shutova E, Pilehvar MT (Association for Com-

putational Linguistics, Vienna, Austria), pp 7583–7604. DOI:10.18653/v1/2025.acl-long.375. Available at <https://aclanthology.org/2025.acl-long.375/>

Appendix A Acronyms

CTA Contextual Token Alignment

GSE Generalized Semantic Embedding

LLM Large Language Model

SSE Sentence-Level Semantic Encoder

Appendix B Pseudo-Code for Synthetic Data Generation

The following algorithm generates a synthetic corpus from a grounded corpus.

B.1 Task Definition

Algorithm 1: Task definition pseudo-code

Define a set of tasks $T = \{t_1, t_2, \dots, t_n\}$ where each task t_i is defined by a meta-prompt m_i .

for each task t_i in T **do**

 Generate a prompt p_i using a meta-LLM based on m_i .

 Store the prompt p_i in a task object.

end for

B.2 Document Processing

Algorithm 2: Document processing pseudo-code

```
for each document  $d$  in the grounded corpus (in batches) do
  for each task  $t_i$  in  $T$  (in parallel) do
    Apply the task  $t_i$  to the document  $d$  using an LLM.
    Store the resulting document  $d'_i$ .
  end for
  Perform cumulative processing on the document  $d$  using the tasks in  $T$ .
  Store the resulting cumulative document  $d'_c$ .
  Validate the cumulative document  $d'_c$  against the original document  $d$  using an LLM.
end for
```

B.3 Corpus Generation

Algorithm 3: Corpus generation pseudo-code

```
for each document  $d$  in the grounded corpus do
  Apply tasks in  $T$  to  $d$  to generate modified documents  $d'_i$ .
  Perform cumulative processing to generate  $d'_c$ .
  Validate  $d'_c$  against  $d$ .
  Store  $d'_i$  and  $d'_c$  in a synthetic corpus.
end for
```

Appendix C Document Examples

C.1 High Similarity Score

C.1.1 Original

Continuing Public Outreach over Time

The outreach strategy should address both the planning process and how to keep people engaged after the plans adoption. Ongoing outreach continues the discussion with the community about hazards and risks, builds support for implementation of mitigation activities, and informs the outreach strategy for the next plan update process. The plan must describe how the jurisdictions will continue public participation during the plans implementation and maintenance. The outreach activities conducted during the planning process, as described above, are a good source of ideas for how to continue to

involve stakeholders and the public during plan maintenance and implementation. Consider repeating successful outreach events annually. Other examples of activities for continued public participation include periodic presentations on the plans progress to elected officials, schools, or other community groups; annual questionnaires or surveys; postings on social media and email lists; and interactive websites. You may help build capabilities throughout the planning area by assigning the responsibility for coordinating these activities to a staff member in each jurisdiction.

Coordinating a Multi-Jurisdictional Outreach Strategy

If you are developing a multi-jurisdictional plan, the outreach strategy creates a mechanism for coordination and accountability among the jurisdictions. For each jurisdiction seeking plan approval, the plan must document how they were involved in the planning process, including how they provided opportunities for the involvement of their stakeholders and the public.

Task 2 describes ways that the representatives on multi-jurisdictional planning teams can share information with their respective community stakeholders and citizens. Specific stakeholders may be identified for each participating jurisdiction, and public involvement activities should be designed to reach citizens throughout the planning area. The planning team may develop one set of outreach materials, which each jurisdiction is responsible for distributing or hosting to reach their stakeholders and citizens. Another good approach is to develop one presentation, or a series of presentations on the plans progress, that jurisdictional representatives can deliver at a regularly scheduled open meeting of their city council or governing body. This is a good method to keep elected officials informed of the planning project and to give the public an opportunity to be informed and provide comments.

C.1.2 Content Augmentation Modification (GSE Score: 0.995)

Continuing Public Outreach over Time

The outreach strategy should address both the planning process and how to keep people engaged after the plan's adoption. Ongoing outreach continues the discussion with the community about hazards and risks, builds support for implementation of mitigation activities (actions taken to reduce or eliminate risks), and informs the outreach strategy for the next plan update process. The plan must describe how the jurisdictions (local governments or organizations responsible for implementing the plan) will continue public participation during the plan's implementation and maintenance.

The outreach activities conducted during the planning process, as described above, are a good source of ideas for how to continue to involve stakeholders (individuals or groups with an interest in the plan) and the public during plan maintenance and implementation. Consider repeating successful outreach events annually. Other examples of activities for continued public participation include:

- Periodic presentations on the plan's progress to elected officials, schools, or other community groups
- Annual questionnaires or surveys to gather feedback from the public
- Postings on social media and email lists to keep the public informed
- Interactive websites that allow the public to access information and provide comments

You may help build capabilities throughout the planning area by assigning the responsibility for coordinating these activities to a staff member in each jurisdiction. This can help ensure that outreach efforts are sustained over time.

Coordinating a Multi-Jurisdictional Outreach Strategy

If you are developing a multi-jurisdictional plan (a plan that involves multiple local governments or organizations), the outreach strategy creates a mechanism for coordination and accountability among the jurisdictions. For each jurisdiction seeking plan approval, the plan must document how they were involved in the planning process, including how they provided opportunities for the involvement of their stakeholders and the public.

Task 2 describes ways that the representatives on multi-jurisdictional planning teams can share information with their respective community stakeholders and citizens. Specific stakeholders may be identified for each participating jurisdiction, and public involvement activities should be designed to reach citizens throughout the planning area.

The planning team may develop one set of outreach materials, which each jurisdiction is responsible for distributing or hosting to reach their stakeholders and citizens. Another good approach is to develop one presentation, or a series of presentations on the plan's progress, that jurisdictional representatives can deliver at a regularly scheduled open meeting of their city council or governing body. This is a good method to keep elected officials informed of the planning project and to give the public an opportunity to be informed and provide comments.

C.2 Low Similarity Score (Hallucination)

C.2.1 Original

Workshop for rapid collection of responses to qualitative questions only (suggested methodology further)

C.2.2 Cumulative Modification (CTA: 0.377)

Unlocking Stakeholder Insights: A Guide to Effective Engagement

Understanding the thoughts and opinions of stakeholders is a valuable way to inform decision-making and drive change. This guide outlines a straightforward process for gathering feedback and opinions from stakeholders.

The Objective of Stakeholder Engagement

The goal is to gather stakeholders' thoughts and opinions on a specific topic by asking relevant questions. This helps to understand their needs, concerns, and expectations, ultimately informing decisions and actions.

The Stakeholder Engagement Process

To achieve this, follow these steps:

1. **Identify and Invite Stakeholders:** Reach out to individuals with a vested interest in the topic or issue. Ensure they understand the purpose of the engagement and are comfortable sharing their opinions. Confidentiality can be assured if needed.
2. **Prepare Open-Ended Questions:** Develop a list of questions that encourage stakeholders to share their thoughts and feelings in detail. Use questions that start with "What," "How," or "Can you tell me more about..." to gather comprehensive feedback.
3. **Create a Suitable Setting:** Choose a setting where stakeholders feel comfortable sharing their thoughts, either in person or online. Clearly explain the questions and allow participants sufficient time to provide their input.
4. **Collect and Analyze Insights:** Gather the feedback from stakeholders and review it to identify common themes and patterns. This analysis will help uncover the insights that inform decision-making.
5. **Inform Decision-Making:** Use the insights gathered to inform decisions and drive change.

Best Practices for Stakeholder Engagement

- Consider keeping the group small (around 5-10 people) to facilitate meaningful contributions.
- Foster an environment where stakeholders feel comfortable sharing their opinions without fear of judgment.
- Be transparent about the purpose and process of the engagement, and provide stakeholders with the option to opt-out if desired.

By following these steps and best practices, stakeholders' valuable insights can be gathered, and informed decisions can be made to meet their needs and expectations. Effective stakeholder engagement is essential for driving positive change.

C.3 Failure to Generate a Response

C.3.1 Original

Enter responses in the City result field Enter the requested information as a number in the relevant field, ensuring that it is in the correct units, e.g. entering 99 instead of 99% if the expected response is percentage of households

C.3.2 Jargon Modification (SSE Score: 0.073)

I'm ready to help. What's the document you need me to revise?