

# Human Impersonation Using Large Language Models

Ismail Ben seddik  
isbens@iu.edu

## Abstract

This work investigates impersonation capabilities of Large Language Models (LLMs). It examines how effectively LLMs can mimic an individual’s writing style and opinions based on limited examples of their responses. Using three distinct impersonation methods—zero-shot prompting, few-shot prompting, and retrieval-augmented generation—the performance is evaluated through automated metrics and LLM-based detection. The findings reveal concerning implications for digital identity security and highlight the need for robust detection mechanisms to mitigate impersonation risks in online communications.

## 1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in generating human-like text, raising significant concerns about their potential for misuse through impersonation. With the proliferation of accessible and powerful LLMs, it has become increasingly easy for malicious actors to mimic individuals by replicating their writing style, opinions, and personal characteristics. This capability poses serious risks across various domains, from generating deceptive emails for phishing attacks to creating fake social media posts that appear authentic.

The ease with which these models can be prompted to adopt specific personas has lowered the technical barriers to sophisticated impersonation attacks, making them accessible even to those without specialized knowledge in natural language processing or artificial intelligence. The potential for harm extends beyond individual privacy violations to broader societal impacts, including undermining trust in digital communications, facilitating identity theft, enabling social engineering attacks, and contributing to the spread of misinformation.

As these models become more sophisticated and widely available, the line between authentic

human-generated content and AI-generated impersonations continues to blur, creating an urgent need to understand the capabilities and limitations of different impersonation methods. This study aims to systematically evaluate three distinct approaches to human impersonation using LLMs: zero-shot prompting, few-shot prompting, and retrieval-augmented generation (RAG).

The primary goals of this research are to assess the effectiveness of different LLM-based impersonation methods in mimicking an individual’s writing style and opinions, evaluate the generalizability of these methods across different individuals, and identify potential vulnerabilities and strengths in current impersonation techniques.

To achieve these goals, experiments are conducted on two datasets: one containing responses from a teammate and another with personal responses from the author, myself. By comparing the performance of each method across multiple dimensions—including style similarity, content consistency, and the ability to fool detection systems—we seek to provide a comprehensive understanding of the current state of LLM-based impersonation capabilities.

Through this investigation, we aim to contribute to the ongoing efforts to develop robust detection mechanisms and appropriate safeguards against the misuse of LLMs for impersonation, while also advancing our understanding of the ethical implications of these powerful technologies.

## 2 Dataset & Methods

### 2.1 Dataset

The dataset comprises 50 questions carefully designed to capture authentic expression styles, tones, opinions, and language structures. The questions were deliberately crafted to be open-ended, allowing the respondent to express complex impressions and attitudes in his own words. This approach pro-

vides deeper insights into authentic opinions and language use patterns that are essential for effective impersonation (Hansen and Świdorska, 2024). To create a diverse and representative dataset, we compiled a list of open-ended personal questions from multiple online sources. The selection process prioritized questions that would elicit detailed responses across various topics, including personal preferences, opinions on social issues, and reflections on experiences. This diversity ensures that the impersonation models are trained on a broad spectrum of linguistic patterns and opinion structures. The responses collected were authentic, detailed, and well-thought-out, providing rich linguistic material for the impersonation task. Each answer reflected the respondent’s natural writing style, vocabulary choices, sentence structures, and opinion formation patterns. This authenticity is crucial for training models that can convincingly mimic a specific individual’s communication patterns. Initially, the dataset was split into 60% for training, 20% for validation, and 20% for testing. However, as the investigation progressed, it turned out that the validation set was not required for the specific impersonation methods employed in this study. Consequently, the data allocation was adjusted to 80% for training and 20% for testing, maximizing the amount of data available for training while maintaining a sufficient test set to evaluate performance. The final dataset structure consists of 40 training examples and 10 test examples for each individual (the teammate and the author). This configuration provides sufficient data for the models to learn distinctive linguistic patterns while reserving independent examples for evaluation. The test questions were selected to represent diverse topics, ensuring a comprehensive assessment of the impersonation methods’ effectiveness across different contexts and subject matters.

## 2.2 Language Model

Llama 3 with 7B parameters was used for the impersonation tasks, a state-of-the-art open-source large language model developed by Meta AI. This model represents a significant advancement in the capabilities of relatively compact language models, offering a balance between performance and computational efficiency (Grattafiori et al., 2024). Given the limited compute resources available for this research, the 7B parameter variant was selected as it provides robust language generation capabilities while remaining accessible on consumer-grade

hardware.

Llama 3 builds upon its predecessors with improved instruction-following capabilities, enhanced reasoning, and reduced hallucination rates. The model was trained on a diverse corpus of text data, enabling it to generate coherent and contextually appropriate responses across a wide range of topics. This makes it particularly suitable for impersonation tasks, which require the model to adapt its output to match specific linguistic patterns and opinion structures.

The model was used with different prompting strategies—zero-shot, few-shot, and retrieval-augmented generation—to evaluate which approach yields the most convincing impersonations. For all experiments, we used the default sampling parameters provided by Ollama, with a temperature of 0.8 to balance creativity and coherence in the generated responses.

## 2.3 Impersonation Methods

### 2.3.1 Zero-shot Prompting

Zero-shot prompting is a technique where the model generates responses without being provided with explicit examples of the target task (Chen and Moscholios, 2024). In the implementation, this method did not directly utilize the dataset for training. Instead, data was manually analyzed to identify distinctive language patterns, stylistic elements, and general opinions of the person to be impersonated. These characteristics were then summarized into a comprehensive prompt that instructed the model on how to respond.

Zero-shot prompting relies on the model’s understanding of language, context, and general knowledge acquired during its training phase. The model attempts to generate answers to questions from the test set by drawing on its pre-trained knowledge while adhering to the stylistic guidelines provided in the prompt. This approach is particularly efficient as it eliminates the need to process example inputs for each query.

### 2.3.2 Few-shot Prompting

Few-shot prompting involves providing the language model with a small number of examples to guide its response to a specific task (Brown et al., 2020). For the impersonation task, this method sampled 15 examples from the training set to demonstrate the person’s authentic responses. The number 15 was determined through experimentation as sufficient context for the model to generate

convincing impersonations.

The model analyzes the provided examples to identify patterns in how the person expresses themselves. The model then attempts to generalize from these examples to new, unseen inputs in the test set. Few-shot prompting enables the model to adapt to specific tasks without requiring changes to its underlying parameters.

### 2.3.3 Retrieval-Augmented Generation (RAG)

Retrieval-augmented generation combines the text-generation capabilities of large language models with information retrieval techniques (Gao et al., 2024). The implementation queries a vector database containing the training answers to find the most relevant examples for each test question. For each query, the system retrieves the top three most relevant examples, with relevance measured using cosine similarity between the vector representations of questions.

Using RAG the model responds to user queries with reference to a specified set of documents. This approach allows the model to access domain-specific information—in this case, the person’s authentic responses—that might not be present in its training data. By incorporating these retrieved examples into the prompt, the model can generate responses that more accurately reflect the person’s unique style and opinions.

The RAG approach addresses limitations of both zero-shot and few-shot methods by dynamically selecting the most contextually relevant examples for each specific question, rather than using either no examples or a fixed set of examples.

## 2.4 Evaluation

To comprehensively assess the quality of impersonation across different methods, three distinct evaluation approaches were employed. Each method was selected to capture different aspects of impersonation quality, from stylistic features to semantic content and practical deception capability. This comprehensive assessment framework allows to identify not only which methods produce the most convincing impersonations overall but also which aspects of impersonation each method excels at.

### 2.4.1 Style Similarity

The first evaluation method measures stylistic similarity between generated responses and authentic writing samples. This approach is grounded in stylistometric analysis techniques that have been widely

used in authorship attribution research (Stamatatos, 2009). A composite style similarity score is calculated based on three key linguistic features:

- Average sentence length: captures the typical syntactic complexity preferred by the individual, as sentence length is often a consistent stylistic choice
- Vocabulary richness: the ratio of unique words to total words, this measure reflects lexical diversity and word choice patterns that are distinctive to individual writers.
- Punctuation usage patterns: the frequency and distribution of punctuation marks often reveal unconscious writing habits that can be distinctive identifiers.

These features were selected based on their demonstrated effectiveness in computational stylistics research and their resistance to conscious manipulation (Cer et al., 2018). The style similarity score provides a quantitative measure of how closely the generated text mimics the writing characteristics of the target individual.

### 2.4.2 Content Consistency

Content consistency evaluates semantic alignment between generated responses and authentic opinions. This method employs embedding-based similarity measures to assess how well the generated content preserves the individual’s viewpoints and knowledge (Reimers and Gurevych, 2019).

Sentence transformers were used to encode both the generated answers and the individual’s previous responses into dense vector representations. Cosine similarity between these vectors provides a measure of semantic proximity, indicating how well the generated content aligns with the individual’s established opinions and typical responses to similar topics (Cer et al., 2018).

This approach is particularly important for impersonation assessment, as convincing impersonation requires not only mimicking writing style but also maintaining consistency with the target’s perspectives and beliefs (Krishna et al., 2023).

### 2.4.3 LLM Judge Evaluation

The third evaluation method employs a third-party LLM (Llama 3) as an impartial judge to assess impersonation quality from a detection perspective.

The LLM judge is first exposed to all training data to establish a baseline understanding of the

individual's authentic writing style and opinions. It then evaluates each test sample, attempting to determine whether it was generated by an LLM or written by the actual person. For each evaluation, the judge provides:

1. A binary classification (real or generated)
2. A confidence score (1-10) indicating certainty in the classification

From these judgments, a "fooling rate" is calculated for each impersonation method as the percentage of generated responses that the judge incorrectly classified as authentic. This metric directly measures practical impersonation success, as it quantifies how often the generated text can deceive an AI system trained to detect such impersonations.

### 3 Results & Discussion

The experiments assessed each impersonation method's performance using three evaluation metrics: style similarity, content consistency, and fooling rate (LLM judge).

#### 3.1 Key Findings

Table 1 summarizes the performance of each impersonation method averaged across the test set with my teammate's answers, highlighting the composite score that combines all evaluation metrics.

##### 3.1.1 Generalizability

This study includes assessing the generalizability of impersonation methods across different individuals. To achieve this, parallel experiments using both the teammate's dataset and my own dataset were conducted, each containing 50 questions with authentic responses. This cross-individual evaluation provides valuable insights into whether impersonation methods optimized for one person can effectively transfer to another.

##### 3.1.2 Results Analysis

Figure 1 illustrates the results of the three evaluation metrics across different impersonation methods, categorized by dataset—my own and my teammate's.

For the teammate's dataset, few-shot prompting demonstrated the strongest performance, with the highest composite score (0.544) and fooling rate (0.6). Interestingly, zero-shot prompting achieved

the highest content consistency (0.476) despite having no access to training examples, suggesting that the manual analysis and prompt engineering effectively captured the teammate's opinion patterns.

For my dataset, RAG emerged as the superior method overall, achieving the highest style similarity (0.604) and fooling rate (0.6). This suggests that RAG's ability to dynamically retrieve relevant examples for each query provides a significant advantage in capturing personal writing characteristics.

When considering both datasets together, RAG emerged as the overall best impersonation method, which reveals its effectiveness in knowledge-intensive tasks. This suggests that the dynamic retrieval of relevant examples provides a more adaptable approach to impersonation across different individuals.

**Zero-shot prompting** performed surprisingly well on content consistency for the teammate's dataset, outperforming both few-shot and RAG approaches. This unexpected result may be attributed to the thorough manual analysis conducted to create the prompt, which effectively captured the teammate's core opinions and perspectives. However, zero-shot prompting consistently underperformed in style similarity and fooling rate, indicating that while it can capture general opinions, it struggles to replicate the nuanced stylistic elements that make writing recognizably authentic.

**Few-shot prompting** demonstrated strong performance across both datasets, particularly excelling with the teammate's data. The method's effectiveness can be attributed to its ability to provide the model with diverse examples that showcase various aspects of the individual's writing style and opinions. The selection of 15 examples appears to have struck an effective balance between providing sufficient context and avoiding information overload.

**RAG** showed the most consistent performance across datasets, achieving the highest overall composite score. Its particular strength in style similarity for my dataset suggests that the method excels at capturing and reproducing distinctive writing patterns when provided with relevant examples. The method's effectiveness in fooling the LLM judge (0.6 fooling rate for my dataset) indicates its practical utility for convincing impersonation.

#### 3.2 Challenges & Future Work

Challenges and limitations encountered include:

Dataset	Method	Style Similarity	Content Consistency	Fooling Rate	Composite Score
teammate	Few-shot	0.4925	0.465	0.5	0.4887
teammate	RAG	0.4349	0.467	0.5	0.4753
teammate	Zero-shot	0.4456	0.476	0.4	0.4303

Table 1: Average Performance of Impersonation Methods on the Teammate’s Dataset.

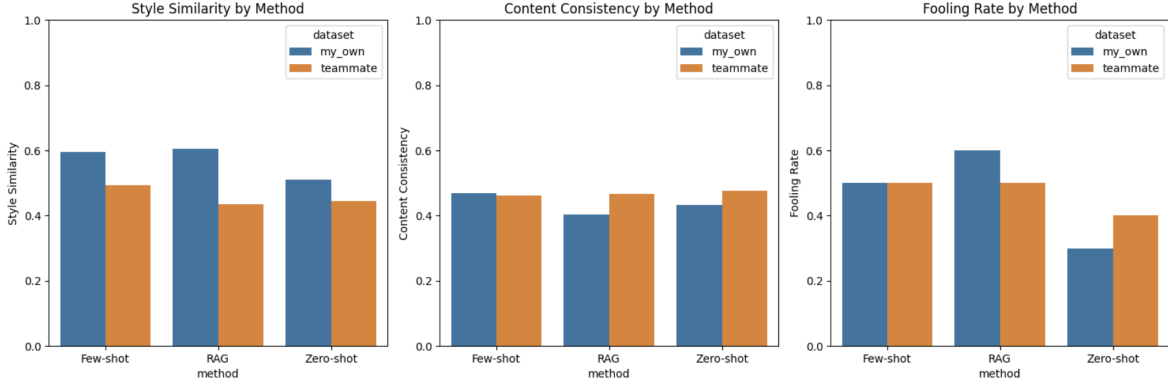


Figure 1: Three Evaluation Metrics Across Different Impersonation Methods.

- **Computational constraints:** The use of a 7B parameter model, while practical for local execution, limited the potential performance of all methods. Larger models with expanded context windows would likely yield improved results, particularly for few-shot prompting where more examples could be included.
- **Dataset size:** The relatively small dataset (50 questions) constrained the diversity of examples available for training and evaluation. This limitation particularly affected RAG, which relies on finding relevant examples in the knowledge base. As noted by (Guu et al., 2020), larger retrieval databases typically improve RAG performance by increasing the likelihood of finding closely matched examples.
- **Fine-tuning limitations:** The computational resources available precluded the exploration of fine-tuning approaches, which have shown promise in personalization tasks. Even with a 7B parameter model, fine-tuning requires substantial computational resources beyond what was available.

Future work would include:

- Employing more powerful models like GPT-4 and curating larger, more diverse datasets would likely yield improved impersonation capabilities and more robust evaluation.

- Exploring chain-of-thought prompting and other advanced prompting strategies could enhance the model’s ability to reason about and reproduce an individual’s distinctive reasoning patterns.
- Testing impersonation methods across different types of writing (formal, informal, technical, creative) would provide insights into the generalizability of these approaches.

#### 4 Ethical Considerations

LLM-based impersonation poses substantial privacy risks through multiple vectors. These models can memorize and reproduce sensitive information from their training data, potentially exposing personal details in generated content (Sentra, 2025). More concerning is their ability to be weaponized for identity theft, where bad actors can train models to mimic a person’s writing style and generate messages that appear to come from that individual. This capability extends to financial fraud, where LLM bots can be primed with communication patterns to impersonate individuals when interacting with financial institutions.

The security implications of LLM-based impersonation are far-reaching. Cybercriminals can exploit these technologies to conduct sophisticated social engineering attacks, tricking users into revealing sensitive information or clicking on malicious links. These attacks are becoming increasingly difficult to detect as the technology advances.

Beyond individual harm, LLM-based impersonation threatens broader societal structures. The technology can be used to spread misinformation and disinformation by creating realistic but false content attributed to public figures, potentially manipulating public opinion, influencing elections, and destabilizing societies.

As impersonation technologies become more sophisticated, distinguishing between authentic and fake content becomes increasingly difficult. This erosion of trust can have far-reaching consequences, undermining public confidence in news organizations, social media platforms, and digital content as a whole. The potential for AI-generated impersonations to mimic public figures such as politicians, celebrities, or influencers creates opportunities for widespread deception that can have significant societal and financial consequences.

Addressing these ethical challenges requires a multi-faceted approach. Developers must adhere to strict ethical standards, including obtaining proper consent, ensuring transparency in AI development, and implementing robust accountability mechanisms. On a technical level, mitigating these risks requires training LLMs with diverse datasets to enhance fairness and minimize biases. Regular ethical audits and consultations with legal experts are essential to ensure compliance with evolving standards. For individuals, implementing protective measures such as safe words or secret codes that only family members know can help verify identity in suspicious communications.

## 5 AI Use Declaration

ChatGPT was used to generate some of the questions included in the dataset. It served as a tool to diversify the question pool and ensure a comprehensive range of topics that could elicit distinctive responses from participants. Additionally, I employed ChatGPT 3.5 to explore and refine different evaluation methodologies before implementing them in the final design.

For the implementation phase, I utilized ChatGPT 3.5 to assist with certain code sections where it provided more accessible explanations than available documentation or online resources. This primarily involved code for the evaluation metrics and running the experiment.

Regarding the report itself, AI assistance was limited to polishing the writing and fixing grammatical errors. All research design decisions, analysis

of results, and conclusions drawn were conducted independently. The core intellectual contribution, including the experimental design, interpretation of findings, and ethical considerations, remains my original work.

This declaration ensures transparency about the role of AI tools in this research while maintaining the academic integrity of the work presented.

## References

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.
- Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, and 1 others. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.
- Ziyang Chen and Stylios Moscholios. 2024. [Using prompts to guide large language models in imitating a real person's language style](#). *Preprint*, arXiv:2410.03848.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. [Retrieval-augmented generation for large language models: A survey](#). *Preprint*, arXiv:2312.10997.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Papatat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.
- Karolina Hansen and Aleksandra Świdarska. 2024. Integrating open- and closed-ended questions on attitudes towards outgroups with different methods of text analysis. *Behav. Res. Methods*, 56(5):4802–4822.
- Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. [Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense](#). *Preprint*, arXiv:2303.13408.

Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). *Preprint*, arXiv:1908.10084.

Sentra. 2025. Safeguarding data integrity and privacy in the age of ai-powered large language models (llms). Accessed: March 31, 2025.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.*, 60(3):538–556.

## A Appendix

### A.1 Dataset Questions

1. What do you do for fun on weekends?
2. What was the last book you read, and what did you think about it?
3. Do you agree with the statement: "AI will replace most human jobs"? Why or why not?
4. Do you play any games? If so, which ones and why? If not, why not?
5. What's the most recent movie you watched, and what was your opinion?
6. Do you use X (formerly Twitter)? If so, for what purpose?
7. What is your religion (if any), and what are its core values? If none, why not?
8. What are your thoughts on climate change?
9. How do you feel about the increasing use of AI in law enforcement?
10. Do you believe privacy is more important than security, or vice versa? Explain your stance.
11. What is your favorite way to spend a day off?
12. If you could have dinner with any historical figure, who would it be and why?
13. What motivates you to keep learning and growing?
14. How do you typically approach problem-solving?
15. What role does music play in your life?
16. How do you usually make difficult decisions?
17. What is one experience that has significantly shaped your worldview?
18. How would you describe your communication style?
19. What values are most important to you in friendships?
20. How do you define success?
21. What is one habit you have that you would like to change or improve?
22. If you could master any skill instantly, what would it be?
23. How do you usually react to unexpected challenges?

24. What is a controversial opinion you hold?
25. How do you usually express creativity in your daily life?
26. What is a cause or movement that you strongly support?
27. How do you prefer to receive feedback?
28. What personal trait do you think is most misunderstood by others?
29. How do you handle conflicts in personal or professional relationships?
30. What is your favorite childhood memory?
31. If you could relive any moment in your life, what would it be and why?
32. What do you think is the most important quality in a leader?
33. How do you balance personal and professional life?
34. What is one thing you believe people should do more often?
35. How do you usually prepare for big decisions or changes in your life?
36. If you could change one thing about modern society, what would it be?
37. What role do traditions play in your life?
38. How do you stay motivated during difficult times?
39. If you could live in any time period, past or future, which would you choose and why?
40. What is one misconception people have about you?
41. What are your thoughts on the importance of lifelong learning?
42. How do you approach goal-setting in your life?
43. What is the most valuable lesson you've learned from a failure?
44. How do you usually approach making new connections or friendships?
45. What is a small act of kindness that has had a big impact on you?
46. How do you handle situations where your beliefs are challenged?
47. What do you think is the key to happiness?
48. If you could give one piece of advice to your younger self, what would it be?
49. How do you define personal growth?
50. What is your philosophy on work-life balance?

### A.2 Answer Generation

Table 2 presents a comparison of the generation methods' responses to the question 'How do you

cope with failure?’ alongside the actual answer to the question.

Method	Generated Answer	True Answer
Zero-shot	Failure is an inevitable part of growth, especially when exploring uncharted territories like AI and ML. For me, coping with failure means acknowledging it as a stepping stone for learning and improvement. I take time to reflect on what went wrong, and then apply the insights gained to refine my approach, much like the ancient Hindu concept of "purity through imperfection."	I see failure as a learning experience, even though it can be frustrating. I try to analyze what went wrong, improve, and move forward. It's not always easy, but staying resilient and adapting is important.
Few-shot	I cope with failure by reflecting on what went wrong and trying to learn from it. I believe that every mistake is an opportunity for growth, as long as we approach it with a willingness to learn and improve. It's essential not to get too attached to success or failure; instead, focus on the process and how you can refine it.	I see failure as a learning experience, even though it can be frustrating. I try to analyze what went wrong, improve, and move forward. It's not always easy, but staying resilient and adapting is important.
RAG	I believe that everyone makes mistakes, and it's how we learn from them that matters. When I face failure, I try to reflect on what went wrong and what I can improve on next time. It's not about dwelling on the past or getting discouraged, but rather using those experiences as stepping stones for growth. I also think it's essential to surround yourself with people who support and encourage you, even when things don't go as planned. By doing so, you'll find that failure becomes an opportunity to learn and become a better version of yourself.	I see failure as a learning experience, even though it can be frustrating. I try to analyze what went wrong, improve, and move forward. It's not always easy, but staying resilient and adapting is important.

Table 2: Comparison of methods with generated and true answers.