Love from Machine: Personalized Blessing Generation with Flan-T5

Zihe Ban

University of Michigan

Ann Arbor, MI 48104 USA

***Abstract* — We explore how prompt specificity and fine-tuning impact greeting text generation using Flan-T5. Human evaluation shows fine-tuning improves fluency and relevance, especially for detailed prompts. Results suggest potential for real-world applications in personalized message generation.**

***Index Terms* — text generation, prompt engineering, fine-tuning**

I. INTRODUCTION

A. Background and Project Goal

Blessings fill people's lives every day. Cheerful greetings like "Happy Birthday!", "Happy Anniversary!" or "Good luck on your exams!" are common ways to express warmth and care. Yet, many people still find it challenging to translate their true feelings into thoughtful words. Moreover, during festive seasons, individuals often feel overwhelmed by the desire to warmly greet numerous friends and family members individually.

To address this common challenge, this project aims to create an application leveraging Hugging Face's NLP models to automatically generate personalized and sincere greeting messages. By blending technology and genuine human warmth, this tool will help people effortlessly share genuine kindness, making emotional connections simpler and more meaningful.

B. Literature Review

The model selected for this project is Flan-T5, a state-of-the-art instruction-tuned language model developed by Google Research and available through Hugging Face's transformers library. Chung et al. [1] argued that scaling both model size and the number of instruction-following tasks during training in Flan-T5 led to better generalization on unseen tasks [2]. This makes Flan-T5 a promising candidate for generating responses to prompts that require task-specific understanding such as writing a blessing message.

In addition to instruction tuning, recent work in empathetic and personalized text generation provides strong motivation for this project. Rashkin et al. [3] highlighted the importance of emotional relevance in generated responses. Similarly, some work emphasized the need for emotional grounding and personalization in open-domain dialogue agents [4].

II. METHOD

A. Problem formulation

The goal of this project is to investigate how prompt specificity and model fine-tuning affect the quality of text generation in a greeting message generation task. The problem is formulated as a conditional text generation task, where the input x is a natural language prompt describing the desired blessing, and the output y is a fluent, relevant, and logically coherent message generated by a language model.

B. Dataset Construction

We designed a set of 12 evaluation prompts, evenly divided into four categories: Birthday, Anniversary, Encouragement, and Thank You. Each category contains three prompt levels of increasing specificity:

- Level 1: general message (e.g., *Write a birthday message*),
- Level 2: message with relational context (e.g., *Write a birthday message to your best friend*),
- Level 3: message with relational context and detailed personalization (e.g., *Write a birthday message to your best friend Tom, thanking him for always being there*).

To fine-tune the model, we constructed a custom dataset of 84 prompt–output pairs, also organized by the same categories and specificity levels. The responses were adapted and revised from GPT-4 outputs to ensure clarity, emotional warmth, and structural consistency.

C. Model Setup and Fine-Tuning

We fine-tuned the google/flan-t5-base model using LoRA (Low-Rank Adaptation) to reduce computational cost. The dataset was split into 80% training and 20% validation sets. Training was conducted for up to 40 epochs with early stopping based on validation loss, using a batch size of 4 and a learning rate of $1e-4$. Evaluation loss and generation quality were monitored using Hugging Face's Seq2SeqTrainer API with an early stopping patience of 2 epochs.

The final model was saved and used to generate messages from the same set of 12 evaluation prompts for comparison with the base model.

D. *Evaluation Procedure*

To assess output quality, we manually rated each generated message using three human-judged dimensions, on a scale from 1 (poor) to 5 (excellent):

- Fluency: Whether the output is grammatically correct and well-formed.
- Relevance: Whether the message addresses the key elements in the prompt (e.g., occasion, relationship, or details).
- Logic: Whether the output is coherent and reads like a real, complete greeting.

This evaluation was applied to messages generated both before and after fine-tuning, across all prompt levels. While BLEU and other automated metrics are commonly used for language generation tasks, they were not adopted in this project. Given the short and stylistically varied nature of blessing messages, such metrics often fail to capture nuance, emotional tone, and personalization.

III. RESULTS

Generated outputs and all evaluation scores are included in the project GitHub repository, for full reproducibility and inspection.

A. *Data Pipeline and Model Execution*

The complete model pipeline involved prompt formatting, text generation using both the base and fine-tuned models, and manual scoring of outputs. Generated messages were stored in structured CSV files and analyzed in Python using pandas, matplotlib, and seaborn.

B. *Performance Comparison Across Prompt Levels*

We evaluated how varying prompt specificity influences output quality in both the base model (Flan-T5) and the fine-tuned model, as shown in Figure 1.

In the base model, increased prompt specificity led to moderate changes in relevance, logic and fluency, suggesting limitations in the model's ability to leverage fine-grained input.

In contrast, the fine-tuned model exhibited a more pronounced improvement across levels. Level 3 prompts resulted in consistently higher scores in relevance and logic, indicating the model's enhanced ability to incorporate nuanced input.

Overall, both models benefited from richer prompts more or less, but the fine-tuned model was more responsive to layered prompt structures. This validates our hypothesis that prompt engineering, when combined with task-specific fine-tuning, can significantly enhance the alignment between input intent and model output.

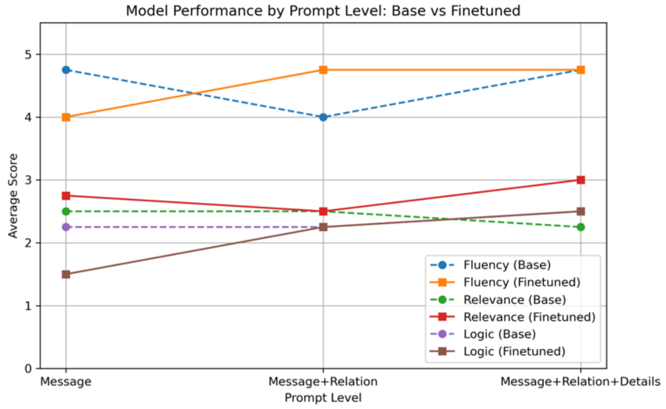

Fig. 1. Model Performance by Prompt Level

C. *Impact of Fine-Tuning*

To assess the effectiveness of model fine-tuning, we compared the average dimension scores of the fine-tuned model against the base model across all prompt levels. The resulting heatmap (Figure 2) illustrates the score improvements (or degradations) for fluency, relevance, and logic at each level.

The most notable gains occurred in Level 2 and Level 3 prompts. For Level 2, fluency improved by 0.75 points, while Level 3 showed a 0.75-point gain in relevance. These results suggest that fine-tuning enhanced the model's ability to generate more fluent and relevant responses when relational and personalized context were introduced.

In contrast, Level 1 performance declined in fluency and logic, each dropping by 0.75 points. This degradation may reflect a trade-off introduced during fine-tuning: the model became more attuned to handling detailed prompts but less optimized for generic ones.

Overall, fine-tuning effectively boosted the model's capacity to handle relational and personalized instructions, aligning better with the structure and intent of richer prompts. However, the loss of performance in simpler prompts underscores the importance of balancing generalization and specialization in instructional fine-tuning.

As an example, given the Level 2 prompt "*Write a warm birthday message to your best friend Tom.*":

- Base model: "*tom! You're welcome! I can see you and your best friend! I love that. I can understand it. But, you're not the only one who has a crush on you. And your heart is always there. And you have a great time. That makes you feel like you've been growing up. That doesn't stop people from thinking about each*"
- Finetuned model: "*Happy birthday! I'm sooo excited! It's a great day. But you are in love with your best friend! You can make it feel like royalty.*"

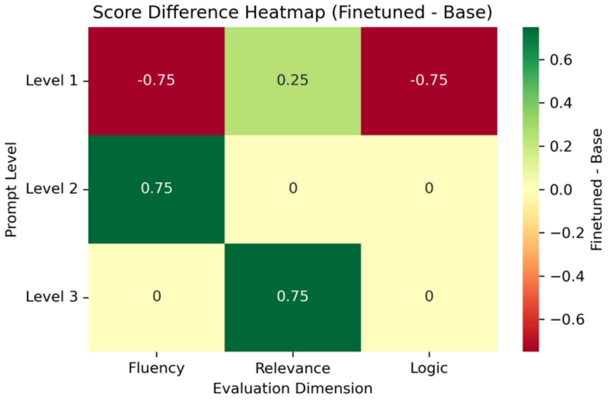

Fig. 2. Score Difference Heatmap (Finetuned - Base)

IV. CONCLUSION

This project examined how prompt specificity and model fine-tuning influence greeting message generation using Flan-T5. Results indicate that the base model responded weakly to prompt variations, while the fine-tuned model demonstrated clear gains in fluency and relevance, particularly with more detailed prompts.

However, the study also faced limitations. The small fine-tuning dataset may have constrained the model's potential, and the human evaluation, though carefully executed, inevitably involved subjective judgments. These factors suggest that future work should incorporate larger, more diverse training data and potentially more objective evaluation metrics.

Despite these constraints, the results are promising. With further development, this system could serve as a practical prototype for applications that help users effortlessly generate thoughtful, personalized messages, making emotional communication easier and more meaningful.

REFERENCES

[1] Chung, Hyung Won et al., "Scaling Instruction-Finetuned Language Models," 2022-10, doi: 10.48550/arxiv.2210.11416.

[2] Wei, Jason et al., "Finetuned Language Models Are Zero-Shot Learners," 2021-09, doi: 10.48550/arxiv.2109.01652.

[3] Rashkin, Hannah, Smith, Eric Michael, Li, Margaret, and Boureau, Y-Lan, "Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset," 2018-10, doi: 10.48550/arxiv.1811.00207.

[4] Shuster, Kurt, Ju, Da, Roller, Stephen, Dinan, Emily, Boureau, Y-Lan, and Weston, Jason, "The Dialogue Dodecathlon: Open-Domain Knowledge and Image Grounded Conversational Agents," 2019-11, doi: 10.48550/arxiv.1911.03768.