Fine-tuning Language Models for Recipe Generation: A Comparative Analysis and Benchmark Study

Anneketh Vij [*], **Changhao Liu**[*], **Rahul Anil Nair**[*], **Theo Ho**[*], **Edward Shi, Ayan Bhowmick**
Department of Computer Science
University of Southern California
Los Angeles, CA 90007
{anneketh, celiu, ranair, teho, epshi, abhowmic}@usc.edu

Abstract

This research presents an exploration and study of the recipe generation task by fine-tuning various very small language models, with a focus on developing robust evaluation metrics and comparing across different language models the open-ended task of recipe generation. This study presents extensive experiments with multiple model architectures, ranging from T5-small (Raffel et al., 2023) and SmolLM-135M (Allal et al., 2024) to Phi-2 (Research, 2023), implementing both traditional NLP metrics and custom domain-specific evaluation metrics. Our novel evaluation framework incorporates recipe-specific metrics for assessing content quality and introduces approaches to allergen substitution. The results indicate that, while larger models generally perform better on standard metrics, the relationship between model size and recipe quality is more nuanced when considering domain-specific metrics. We find that SmolLM-360M and SmolLM-1.7B demonstrate comparable performance despite their size difference, while Phi-2 shows limitations in recipe generation despite its larger parameter count. Our comprehensive evaluation framework and allergen substitution systems provide valuable insights for future work in recipe generation and broader NLG tasks that require domain expertise and safety considerations.

1 Introduction

The generation of safe and high-quality recipes presents unique challenges in natural language generation. Beyond generating coherent and creative recipes, recipe generation requires high-level knowledge of culinary techniques, nutritional principles, and awareness of dietary restrictions to ensure user safety. This necessitates approaches that balance linguistic fluency with domain-specific expertise, particularly in the domain of allergen substitution.

[*]Authors contributed equally.

Our research focuses on addressing these challenges by experimenting with different model architectures for recipe generation and allergen substitution through controlled fine-tuning and comprehensive evaluation metrics. We have focused our research on answering the following three research questions:

1. Given the scope of our research, which models will achieve the best results after fine tuning for recipe generation?

2. How should we evaluate the generated recipes to ensure that they are coherent and safe for users with dietary restrictions?

3. How should we implement allergen substitution into our model to achieve best performance?

To answer these questions, we make the following contributions:

- Comprehensive comparison of model architectures across scales including smaller models like GPT-2 (Radford et al., 2019) and T5 (Raffel et al., 2020) and bigger models like Phi-2 (Research, 2023) and SmolLM-1.7(Allal et al., 2024)

- Multi-dimensional evaluation framework combining novel recipe-specific evaluation metrics, traditional metrics, and LLM-based assessment

- Development of RAG and prompt-based approach for allergen substitution

Our work represents a step forward in adapting NLG systems for practical applications in the culinary domain, emphasizing safety, personalization, and quality.

2 Related Work

2.1 Language Models in Recipe Generation

Previous works have explored various language model architectures for recipe generation. Our work extends this by systematically comparing models across different sizes and architectures, from smaller models like SmolLM-135M to larger ones like Phi-2. We contribute to this area by providing a detailed analysis of how model size impacts different aspects of recipe quality. We have also developed custom evaluation metrics to focus on the impact in these areas.

2.2 Recipe Generation Models

Our work builds on several recent advances in recipe generation and personalization. Majumder et al. (2019) proposed a personalized recipe generation model using attention mechanisms to focus on recipes previously consumed by the user. Their approach showed promising results in generating recipes that aligned with user preferences. We have used the custom encoder-decoder model from this paper as our baseline and obtained our current dataset from it. We have adopted part of the idea of personalization to consider allergen avoidance for specific allergens in our recipes.

Chen et al. (2021) implemented a framework using constrained question answering over a large-scale knowledge graph to recommend food recipes considering users' explicit requirements and health factors. This helped recommend healthy alternatives to users, which aligned with our study's goal of providing allergen-free options.

2.3 Multi-modal Approaches

The FIRE system, by Chhikara et al. (2024) and Nutrify AI by Han and Chen (2024), both use a multi-modal approach, generating recipes from food images and ingredients. While it differs from our core study and exploration due to the fact that we are not using images, it is similar in the idea of incorporating different types of input in recipe generation, since we include specific allergens in our input for substitution.

The LLava-Chef, by Mohbat and Zaki (2024) is another multi-modal approach to recipe generation, which was fine-tuned on both the cross-entropy loss and a novel loss function computed using BLEU and ROUGE scores to ensure that the model generated recipes closer to the ground truth. We adopted our evaluation metrics and the idea of creating custom ones from this paper, as well as what inputs to include for recipe generation. However, we avoided using the novel loss function for fine-tuning our models since penalizing them for not being closer to the ground truth might hinder personalization of generated recipes, which was an important part of the allergen substitution.

2.4 Evaluation

Recent studies, like LLava-Chef, the Fire system, and Retrieval Augmented Recipe Generation by (Liu et al., 2024), utilize conventional metrics such as BLEU, ROUGE, F1-score for ingredient matching to assess recipe quality. Our paper distinguishes itself by employing both general and domain-specific metrics like Ingredient Coverage (Liu et al., 2022) to attain a more profound comprehension of the quality implications across many aspects of the generated recipe since traditional metrics focus more on overlap and thus hinder creativity in generation. We compare performance prior to and following fine-tuning to assess the impact of fine-tuning on large-scale models for a creative task such as recipe development.

Since human evaluation is expensive, we have also adopted the approach of evaluating using LLM-as-a-judge by prompting it similar to the Likert scale, and evaluating the generated recipes on different metrics.

3 Approach

3.1 Food.com Dataset

The Food.com dataset (Majumder et al., 2019) contains more than 180,000 recipes and 700,000 recipe reviews covering 18 years. Each entry includes the recipe name, the list of ingredients, the cooking instructions, nutritional information, and user ratings and reviews. We used the `RAW_recipes` dataset from the Food.com dataset for our research. Our dataset preprocessing pipeline consisted of the following steps:

- Extraction of recipe names, ingredients lists, and cooking instructions
- Standardization of ingredient formats and measurements
- Tokenization and formatting of recipe names, standardized ingredients and instructions
- Creation of input-output pairs for model training

The format of the input is as follows:-

```
<|startoftext|>[Recipe Name]
Ingredients: [Ingredients List]
```

The cooking instructions were used as the target output for our models.

3.2 Exploratory Data Analysis

We conducted a statistical analysis of the entire dataset to gain insights into the distribution of ingredients and recipe length.

The distribution of ingredient occurrences is dominated by a few common ingredients such as salt, butter, sugar, etc. When considering the set of unique ingredients, 9.66% were included in 90% of the recipes, while the remaining 91.44% were only included in 10% of the recipes.

The tokenized length of recipes was also measured. 99. 4% of the recipes had a tokenized length of less than 512 tokens and 90.4% had less than 256. We used these statistics to determine the size of our context when training our models. Additional analysis can be found in Appendix A.

3.3 Fine-Tuning Small Scale Models

From our dataset, we randomly sampled 100,000 recipes. This dataset was then split into training (80%), validation (10%), and test (10%) sets. For our final evaluation, we used the first 500 samples from the test set to ensure consistency across different model evaluations. We initially implemented a custom encoder-decoder model with attention, inspired by the architecture described in Bahdanau et al. (2016). The model consisted of an embedding layer, a bidirectional GRU encoder, a GRU decoder with attention mechanism, and a final linear layer for output generation. However, this model produced near-zero scores on our evaluation metrics, indicating significant challenges in learning the complex patterns required for recipe generation. Following the challenges with the custom model, we turned to pre-trained language models, such as SmolLM (Allal et al., 2024) (135 M), GPT-2(small and medium variants)(Radford et al., 2019), and encoder-decoder language models like T5-small (Raffel et al., 2023) to explore the impact of model size and architecture on recipe generation. We fine-tuned these models on our recipe dataset, using the following approach:

- Input: Combined recipe name and ingredients
- Output: Cooking instructions

Training configurations for the small-scale models are listed in Appendix I, as we experimented with different hyperparameters for different models according to the constraints required for fine-tuning a particular model. A sample output for these small-scale models is given in Appendix B

3.4 Fine-Tuning Larger Models

From the evaluation metrics of the generations of the small-scale models, as seen in Table 1, we decided to scale up the size of our dataset to now include the entire dataset and turned towards large-scale models such as SmolLM-360M, SmolLM-1.7B and Phi-2 instead. We achieved this with our limited computational resources by using the QLORA approach and setting the rank to 8. The entire data set, consisting of 231637 recipes, was split into training (80%), validation (10%), and test (10%) sets. As before, the first 500 samples of the test set were used for evaluation to ensure consistency between the evaluation results for the different model generations. Since these were large-scale models, generation evaluation was performed for both baseline and fine-tuned versions to better understand the impact of fine-tuning on the generated recipes. These models were fine-tuned in the same way as the above models with the training configurations listed in Appendix L. Due to time and resource constraints, all of these models were trained on 1 epoch on these configurations for 8 hours on 2 NVIDIA A100 GPUs.

3.5 Allergen Substitution

Allergen substitution in the generated recipes was conducted using the following two approaches:-

3.5.1 Prompt based Allergen Substitution

Since we had fine-tuned three large-scale models on the entire data set, we hypothesized that these models should be powerful enough to substitute the allergens present in the generated recipe just by prompting. This was done by adding a list of allergens to avoid while generating recipes in the prompt along with the recipe name and the ingredient list. In order to test this approach, we added some common allergens such as milk, eggs, and fish to the list of allergens to avoid and let the models generate accordingly. The prompt is given as follows:-

> "You are an expert chef and recipe writer with a deep understanding of culinary techniques and food allergies. Your goal

is to create a detailed and high quality recipe that uses the provided list of ingredients, while making substitutions for any allergens to ensure the recipe is safe for individuals with those allergies. Please follow these instructions:

1. Create a Recipe: Write a full, detailed recipe based on the name and ingredients provided.
2. Substitute Allergens: Some people are allergic to certain ingredients. You must avoid these allergens in the recipe and suggest substitutions from the list of safe ingredients. If the allergen is an essential part of the recipe, ensure the substitute maintains the flavor and texture as much as possible.
3. Ensure Clarity and Detail: Provide precise instructions, including cooking methods, preparation steps, and any necessary tips. The recipe should be easy to follow for someone with basic cooking knowledge.

Create a recipe for: name
Using these ingredients: ingredients
Substitute these allergens for other ingredients: allergens
Recipe:"

A sample output for these models with and without allergen substitution is given in Appendix D.

3.5.2 RAG-assisted Allergen Substitution System

We also implemented an experimental RAG-assisted allergen substitution system (Lewis et al., 2021) to replace allergens in the generated recipes with similar ingredients as mentioned in a custom allergen database that we built to test the substitution of specific ingredients present in the generated recipes. Key components include:

- FAISS vector store for efficient similarity search
- HuggingFace embeddings (sentence-transformers/all-MiniLM-L6-v2)
- Custom allergen database with substitution rules
- Ingredient parsing and validation system

Implementation details:

- Chunk size: 1000 tokens
- Chunk overlap: 200 tokens
- Top-k retrieval: k=1 for substitution matches

A workflow for the RAG-assisted system can be found in Appendix C. The system finds the ingredients present in the generated recipe and, if they are present in the allergen database, substitutes them with an appropriate ingredient that is mentioned in the database along with a note that states what ingredient is being substituted. This allergen ingredient database can be seen in Appendix F.

4 Evaluation Metrics

A comprehensive evaluation framework has been implemented to evaluate the generation of the recipes of these models. These can be divided into three parts.

4.1 Traditional NLP Metrics

We have implemented traditional NLP metrics in order to evaluate the generated recipes of our models.

1. BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) is a metric that evaluates generated text by comparing it with the ground truth. It evaluates the generated recipe by comparing the n-grams between the generated recipe and the ground truth recipe, and assigning a score between 0 and 1.

2. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004), is also a metric that evaluates the generated recipe by comparing the overlap between the generated recipe and the ground truth. For our research, we have adopted ROUGE 1, ROUGE 2 and ROUGE L for evaluation.

3. Perplexity is another traditional NLP metric that is used to measure the quality of the generated text. It is calculated as the exponentiated average negative log-likelihood of a sequence.

4.2 Recipe Specific Auto Evaluation Metrics

The traditional metrics above are good for measuring overlap with the ground truth. However, they do not work well for evaluating a creative task such as generating recipes. A high quality generated recipe could be given a low score because it does not overlap with the ground truth. Therefore, we have implemented custom auto-evaluation metrics which are tailored to evaluate recipes in various subdomains.These are as follows:

1. Ingredient Coverage Tracking: Measures how effectively the generated recipe utilizes the input ingredients. It tokenizes the ingredient list, matches the ingredients in the generated instructions, and then calculates the coverage ratio, which is the number of present ingredients divided by the total number of ingredients. It can handle several variations and forms.

2. Step Complexity: Evaluates instruction completeness and detail. This is done by counting the distinct operations, analyzing the step length and detail, evaluating the parameter specifications, and then calculating the complexity score.

3. Recipe Coherence: Assesses the logical flow and structure of the recipe. It does so by building a step dependency graph, verifying the logical ordering, checking the temporal consistency, and finally calculating the coherence score.

4. Temperature/Time Specification Checks:- Verifies critical cooking parameters by extracting the numerical values of temperature and time in the generated recipe, validating the ranges per method, checking the completeness, and then calculates the final score.

All of these metrics give evaluation scores in the range of 0 and 1, where the higher the score, the better. A more detailed explanation of these metrics can be found in Appendix E.

4.3 LLM-As-Judge

We have used the LLM-as-judge method to evaluate the recipes generated by both the baseline and fine-tuned versions of the models. We initially used Qwen2.5-1.5B Instruct (Yang et al., 2024) (Team, 2024), but shifted to a much larger model in Qwen2.5-7B (Team, 2024) for more accurate scores when judging the quality of the generated recipes. The recipes are evaluated using six Likert scale categories and are judged on a scale of 1-5. These categories are as follows:-

1. Clarity: Instruction comprehensibility

2. Completeness: Coverage of necessary steps

3. Consistency: Logical flow and coherence

4. Practicality: Feasibility of execution

5. Relevance: Alignment with recipe goals

6. Allergen Safety: Checks if allergen is substituted correctly

5 Results

5.1 Initial Results with Small Scale Models

Table 1 presents our initial results which show our comparison of generated recipes with small-scale models as shown in the table. The evaluation was done with BLEU and ROUGE.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3
Custom Encoder-Decoder	0.10	0.02	0.08	0.05	0.01	0.00
SmolLM (Fine-tuned)	0.22	0.03	0.11	0.15	0.04	0.01
GPT-2 (Small)	0.25	0.05	0.15	0.18	0.07	0.03
GPT-2 Med	0.28	0.06	0.17	0.20	0.08	0.04
GPT-2 Med (Fine-Tuned)	0.33	0.07	0.19	0.25	0.11	0.06
T5-Small (Fine-tuned)	0.13	0.04	0.11	0.00	0.00	0.00

Table 1: Comparison of various small scale models

5.2 Results with Large Scale Models

Table 2 and Table 3 give us the evaluation scores of the baseline and fine-tuned versions of the large-scale models for both traditional NLP metrics and domain-specific auto-evaluation metrics. The models have lower BLEU and ROUGE scores because there is not much overlap with the ground truth, hence the use of domain-specific evaluation metrics.

5.3 Results of Prompt-based Allergy Substitution

Table 4 contains the domain-specific auto-evaluation metrics of the baseline and fine-tuned versions of the large-scale models using prompt-based allergy substitution. Table 5 shows the results of the evaluation conducted by Qwen2.5-7B as a judge for the allergen-substituted recipes generated by the baseline and fine-tuned versions of the models. Evaluation is performed on the first 500 samples of the test set. The radar charts of these results are given in Appendix G.

5.4 Results of RAG-Assisted Allergy Substitution

Table 6 contains the domain-specific auto-evaluation metrics of the baseline and fine-tuned

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU4	Perplexity
SmolLM (360M) - Baseline	0.13	0.01	0.07	0.08	0.02	0.01	0.00	125.2
SmolLM (360M) - Finetuned	0.11	0.01	0.06	0.07	0.01	0.01	0.00	**90.67**
SmolLM (1.7B) - Baseline	0.14	0.01	0.07	0.08	0.02	0.01	0.00	171.07
SmolLM (1.7B) - Finetuned	0.11	0.01	0.05	0.07	0.01	0.00	0.00	**112.13**
Phi-2 - Baseline	0.22	0.03	0.10	0.14	0.05	0.02	0.01	58.74
Phi-2 - Finetuned	0.17	0.01	0.07	0.11	0.03	0.01	0.00	78.9

Table 2: Comparison of Large Scale Models using Traditional Metrics

Model	Ingredient Coverage	Step Complexity	Recipe Coherence	Temp. and Time Spec.
SmolLM (360M) - Baseline	0.21	0.93	0.03	0.10
SmolLM (360M) - Finetuned	0.16	**0.98**	0.02	**0.12**
SmolLM (1.7B) - Baseline	0.29	0.84	0.05	0.11
SmolLM (1.7B) - Finetuned	0.27	**0.97**	0.04	0.03
Phi-2 - Baseline	0.59	0.79	0.08	0.329
Phi-2 - Finetuned	0.30	**0.99**	0.07	0.24

Table 3: Comparison of Large Scale Models Using Domain Specific Metrics

Model	Ingredient Coverage	Step Complexity	Recipe Coherence	Temp. and Time Spec.
SmolLM (360M) - Baseline	0.13	0.74	0.04	0.13
SmolLM (360M) - Finetuned	0.11	**0.92**	0.03	0.09
SmolLM (1.7B) - Baseline	0.15	0.77	0.06	0.13
SmolLM (1.7B) - Finetuned	**0.16**	**0.91**	0.05	0.07
Phi-2 - Baseline	0.30	0.82	0.09	0.20
Phi-2 - Finetuned	0.18	**0.99**	0.08	**0.21**

Table 4: Comparison of Prompt based Allergen Substitution using Domain Specific Metrics

versions of the large-scale models with RAG-assisted allergy substitution. Table 7 contains the results of the evaluation conducted by Qwen2.5-7B as a judge. As before, evaluation for the LLM-as-judge is performed on the first 500 samples of the test set. The radar charts of these Qwen2.5-7B results are given in Appendix H.

6 Discussion

Our comprehensive evaluation across model architectures and scales reveals several profound insights about the intersection of recipe generation and allergen awareness, challenging conventional assumptions about model scaling and domain adaptation.

1. **Model Scaling and Performance:** The progression from smaller to larger architectures revealed unexpected patterns in recipe generation capabilities. The SmolLM-1.7B model showed only marginal improvements over its 360M counterpart in practical aspects such as ingredient coverage (0.27 vs 0.16) and recipe coherence (0.04 vs 0.02). More notably, both models achieved strong performance in temperature and time specifications, with the 360M model even outperforming its larger counterpart (0.12 vs 0.03). This finding challenges the common assumption that larger models inherently perform better in domain-specific tasks, suggesting that architectural efficiency and specialized training may be more crucial than raw parameter count.

2. **Fine-tuning Dynamics:** Our most interesting finding comes from the Phi-2 experiments. Despite its sophisticated architecture, Phi-2 exhibited unexpected behavior post-fine-tuning. While its baseline version achieved high scores in ingredient coverage (0.59) and temperature specification (0.329), the fine-tuned version showed significant degradation across multiple metrics, including domain-specific and LLM-as-judge metrics. While the fine-tuned versions of each model showed remarkable improvement in step complexity, Phi-2 showed degradation in the other three metrics, suggesting its improvement in generating recipes in a complete step-by-step manner is done by trading off semantic relations

within the instructions. This shows that conventional fine-tuning approaches may need a revision for larger models in specialized domains.

3. **Allergen Substitution and Evaluation Framework:** Our prompt-based substitution system revealed complex trade-offs between safety and culinary creativity. The fine-tuned SmolLM models demonstrated promising results in allergen safety (scores of 2.57 and 2.54), although these improvements often came at the cost of recipe coherence, similar to the Phi-2 models. The multi-dimensional evaluation approach proved crucial, revealing significant discrepancies between traditional metrics and practical applicability, as exemplified by Phi-2's metrics in both prompt-based and RAG-asssisted allergen substitution.

4. **Comparison between Prompt-based and RAG-assisted Allergen Substitution Systems :** For the domain-specific metrics, the RAG-assisted method shows higher scores in step complexity and temperature and time specification compared to the prompt-based method, although it has lower scores in ingredient coverage and recipe coherence. This is most likely due to the fact that the RAG-assisted method has more ingredients to substitute and unlike the prompt-based method, is not able to perform coherence checks on the generated recipes. The higher scores in step complexity, temperature, and time specification are most likely due to how the prompt-based approach struggles to generate a step-by-step recipe when allergens are present in the recipe, whereas the RAG-assisted approach only needs to substitute allergens in the generated recipe. We also find that for the LLM-as-judge metric, the prompt-based method outperforms the RAG-assisted method across all models and metrics. This shows that allergen substitutions alone will not produce high-quality recipes, resulting in lower scores.

7 Future Work

Building on our findings, we identify several promising directions to advance recipe generation with allergen-awareness. Our research suggests two main areas for development.

1. The performance degradation observed in larger models during fine-tuning calls for more sophisticated adaptation approaches. Future work should explore constitutional fine-tuning techniques that better preserve model capabilities while adapting to the culinary domain, complemented by specialized pre-training objectives incorporating culinary domain knowledge. We envision a multi-task learning framework that simultaneously optimizes for recipe quality and allergen safety.

2. Future work should explore multiple datasets for fine-tuning as well as focus on better evaluation metrics and increased size of test set for evaluation. Mitigation of inherent bias in LLM-as-a-judge should be explored as well.

3. Our RAG-based (Lewis et al., 2021) allergen substitution system shows considerable promise, but requires further development. Future research should focus on integrating comprehensive domain-specific knowledge bases for more accurate substitutions, with real-time validation mechanisms ensuring substitution safety while maintaining recipe coherence.

8 Conclusion

This work presents a comprehensive exploration of recipe generation and allergen substitution, demonstrating both the possibilities and challenges in developing practical AI systems for culinary applications. Our systematic evaluation across multiple model scales and architectures provides valuable insights into the relationship between model capacity and domain-specific performance. Our results highlight three key findings.

1. The relationship between model size and performance is not strictly linear in creative tasks, as demonstrated by the comparable performance of SmolLM-360M (Allal et al., 2024) and SmolLM-1.7B in many metrics, suggesting that architectural efficiency may be more crucial than model scale or raw parameter count.

2. The challenge of maintaining recipe quality while implementing allergen substitutions requires careful balancing, as shown by our prompt-based substitution results and validated through an LLM-based evaluation.

Model	Clarity	Completeness	Consistency	Practicality	Relevance	Allergen Safety
SmolLM (360M) - Baseline	2.35	2.4	2.26	2.47	3.02	2.26
SmolLM (360M) - Finetuned	**2.46**	**2.6**	2.114	2.28	2.84	**2.57**
SmolLM (1.7B) - Baseline	2.38	2.42	2.26	2.48	3.01	2.29
SmolLM (1.7B) - Finetuned	**2.42**	**2.57**	2.1	2.28	2.96	**2.54**
Phi-2 - Baseline	2.61	2.54	2.48	2.71	3.04	2.46
Phi-2 - Finetuned	2.29	2.24	2.01	2.04	2.32	2.44

Table 5: Comparison of Prompt based Allergen Substitution using Qwen2.5-7b

Model	Ingredient Coverage	Step Complexity	Recipe Coherence	Temp. and Time Spec.
SmolLM (360M) - Baseline	0.11	0.91	0.03	0.12
SmolLM (360M) - Finetuned	0.09	**0.98**	0.02	**0.13**
SmolLM (1.7B) - Baseline	0.13	0.83	0.06	0.16
SmolLM (1.7B) - Finetuned	0.13	**0.97**	0.06	0.04
Phi-2 - Baseline	0.34	0.82	0.08	0.37
Phi-2 - Finetuned	0.16	**0.99**	**0.12**	0.26

Table 6: Comparison of Rag-Assisted Allergen Substitution using Domain Specific Metrics

Model	Clarity	Completeness	Consistency	Practicality	Relevance	Allergen Safety
SmolLM (360M) - Baseline	2.206	2.188	2.065	2.16	2.42	2.172
SmolLM (360M) - Finetuned	2.167	2.112	1.945	1.97	2.211	**2.283**
SmolLM (1.7B) - Baseline	2.250	2.246	2.095	2.188	2.511	2.251
SmolLM (1.7B) - Finetuned	**2.31**	**2.28**	**2.101**	2.125	2.43	**2.413**
Phi-2 - Baseline	2.335	2.342	2.266	2.368	2.503	2.273
Phi-2 - Finetuned	2.146	2.084	1.998	2.061	2.229	2.163

Table 7: Comparison of Rag-Assisted Allergen Substitution using Qwen2.5-7b

Even just substituting similar ingredients in a generated recipe, as proven in RAG-assisted substitution, is insufficient to meet this problem.

3. Our multi-dimensional evaluation framework reveals that traditional NLP metrics alone are insufficient for assessing recipe generation quality, emphasizing the need for domain-specific metrics. The challenges encountered, particularly in fine-tuning larger models and implementing reliable allergen substitutions, establish a strong foundation for future developments in the recipe generation systems.

Ultimately, this work contributes to the broader field of natural language generation by demonstrating that successful recipe generation systems must balance multiple objectives: linguistic coherence, culinary accuracy, and safety considerations. These insights extend beyond recipe generation to inform the development of other domain-specific LM's where safety and expertise are paramount.

9 Limitations

1. **Computation requirements:** One disadvantage of this study is the amount of computational power required to develop high-quality recipes. It took 8 hours to fine-tune each model on two NVIDIA A100s for one epoch. There were insufficient resources for fine-tuning, therefore the models could not be scaled up further.

2. **Using LLM-as Judge for Evaluation:** While we used a large scale model for evaluation and did our best to accurately analyze the output ratings, LLM-as-a-Judge is quite stochastic in terms of generating scores for each recipe.It is also computationally expensive to calculate scores for a big number of recipes.The LLM also has inherent biases to be aware of.

3. **Evaluation on only a part of test set:**Due to the computational constraints, we were only able to generate and evaluate 500 recipes out of above 20000 recipes in the test set. This limited sample size may not fully represent the diversity and complexity of the entire

dataset.Consequently, our results might not generalize to the complete test set, potentially overlooking performance variations across different recipe types or complexity levels.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Leandro von Werra, and Thomas Wolf. 2024. Smollm - blazingly fast and remarkably powerful.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate. *Preprint*, arXiv:1409.0473.

Yu Chen, Ananya Subburathinam, Ching-Hua Chen, and Mohammed J. Zaki. 2021. Personalized food recommendation as constrained question answering over a large-scale food knowledge graph. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21. ACM.

Prateek Chhikara, Dhiraj Chaurasia, Yifan Jiang, Omkar Masur, and Filip Ilievski. 2024. Fire: Food image to recipe generation. *Preprint*, arXiv:2308.14391.

Michelle Han and Junyao Chen. 2024. Nutrifyai: An ai-powered system for real-time food detection, nutritional analysis, and personalized meal recommendations. *Preprint*, arXiv:2408.10532.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Guoshan Liu, Hailong Yin, Bin Zhu, Jingjing Chen, Chong-Wah Ngo, and Yu-Gang Jiang. 2024. Retrieval augmented recipe generation. *arXiv preprint arXiv:2411.08715*.

Yinhong Liu, Yixuan Su, Ehsan Shareghi, and Nigel Collier. 2022. Plug-and-play recipe generation with content planning. *Preprint*, arXiv:2212.05093.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2019. Generating personalized recipes from historical user preferences. *Preprint*, arXiv:1909.00105.

Fnu Mohbat and Mohammed J. Zaki. 2024. Llava-chef: A multi-modal generative model for food recipes. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 1711–1721. ACM.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Colin Raffel, Noam Shazeer, Adam Roberts, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Microsoft Research. 2023. Phi-2: The surprising power of small language models. Accessed: 2024-12-16.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

An Yang et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

A Exploratory Data Analysis

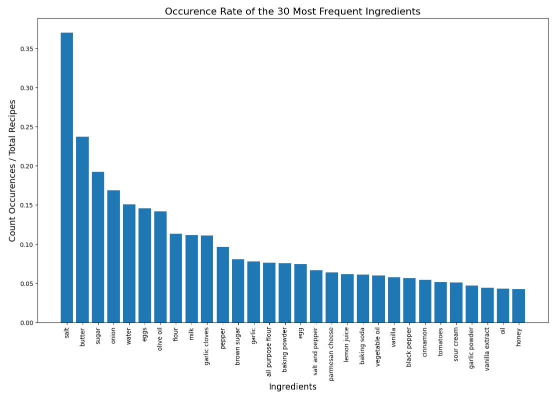

Figure 1: Occurrence rate of the 30 most Frequent Ingredients

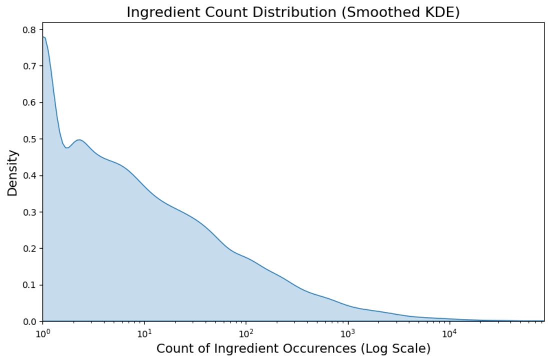

Figure 2: Ingredient Count Distribution (Smoothed KDE)

Figure 3: Distribution of Tokenized Length (Ingredient+Steps)

B Sample Output for Small-scale Model

Section	Content
Input	Chocolate Chip Cookies Ingredients: flour, sugar, butter, chocolate chips, eggs, vanilla extract, baking soda, salt
Baseline Model Output	Preheat oven to 350°F. In a large bowl, cream together the butter and sugar until light and fluffy. Beat in the eggs one at a time, then stir in the vanilla. Combine the flour, baking soda, and salt; gradually stir into the creamed mixture. Fold in the chocolate chips. Drop by rounded tablespoons onto ungreased baking sheets. Bake for 8 to 10 minutes or until golden brown. Cool on wire racks.
Fine-tuned Model Output	Preheat the oven to 375°F (190°C). Line a baking sheet with parchment paper. In a large bowl, cream together the butter and sugar until light and fluffy, about 3-4 minutes. Beat in the eggs one at a time, then stir in the vanilla extract. In a separate bowl, whisk together the flour, baking soda, and salt. Gradually add the dry ingredients to the butter mixture, mixing until just combined. Fold in the chocolate chips. Drop rounded tablespoons of dough onto the prepared baking sheet, spacing them about 2 inches apart. Bake for 9-11 minutes, or until the edges are golden brown. Allow cookies to cool on the baking sheet for 5 minutes before transferring to a wire rack to cool completely.

Table 8: Comparison of Baseline and Fine-tuned Model Outputs for Chocolate Chip Cookie Recipe

C RAG-assisted Allergen Substitution System Workflow

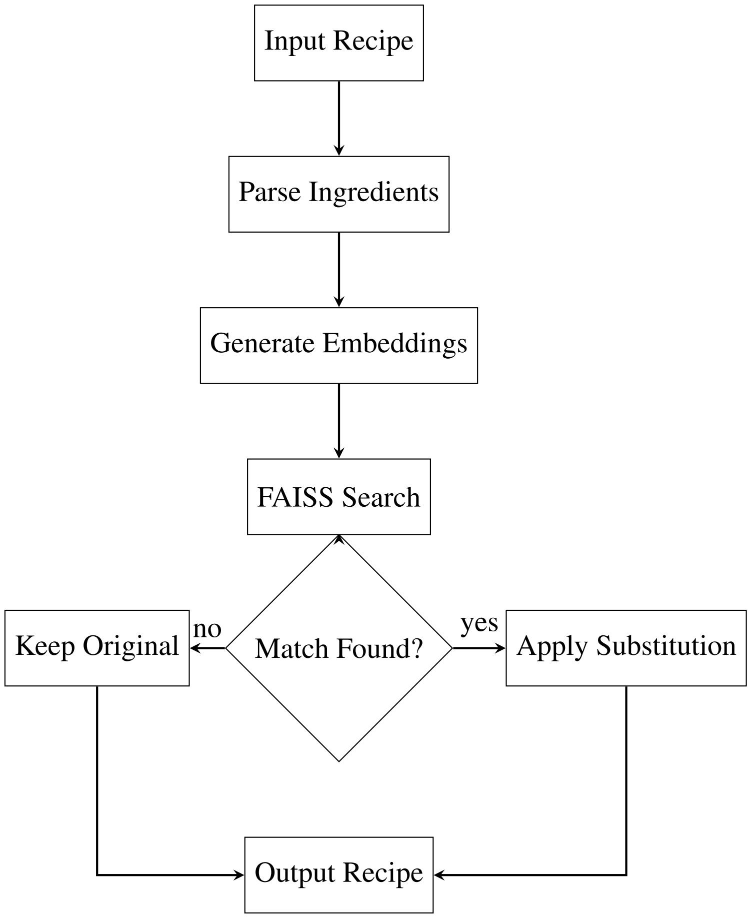

Figure 4: Experimental RAG-based Allergen Substitution System Workflow (Work towards Future Work Component)

D Sample Generated Recipes

The following table shows recipe generations given the following parameters: "name": "cottage cheese waffles", "ingredients": "flour, salt, eggs, milk, vegetable oil, vanilla extract, cottage cheese", and "allergens": ["milk", "eggs", "fish"].

Model	Original Output	With Allergen Substitution
SmolLM-360M:	350 fahrenheit oven cake rolls or cupcakes. 2 cups all purpose whole wheat baking mix caster sugar butter egg white nonfat sour cream almond meal cocoa powder cinnamon pure maple syrup water orange juice lime zest organic dark chocolate nutmeg bourbon whiskey unsweetened condensed sweetened applesauce brown sugar coconut husk powdered sweetener granulated cane soda cornstarch finely ground cardamom peppercorns black pepper dry ice (for frosting) light whipping machine cold kosher buttermilk melted margarine whipped topping blueberry sauce frozen raspberries lemon rind green apple red onion fresh parsley mint jalapeno chili peppers freshly grated pepitas garlic cloves sea bass flakes mayo diced tomatoes chopped cilantro chile relish sweet potato hash chips ham hock bacon seasoning	Soft Serve Waffle Cups with Creamy Cottage Cheese Topping Bake at 400°F (around 256°C) until golden brown around edges & bubbly inside - about 8 minutes or so after turning oven light/heat setting off when ready . Add cream yogurt mix , coconut flakes if desired ! Slice into squares approx., then cut napkins out diagonally just before serving! Enjoy your delicious meal made healthier too :)
SmolLM-1.7B	Cottage Cheese Wafels with Vanilla Ice Cream Topping. 1 cup all purpose flour 2 teaspoons kosher baking powder pinches nutmeg pepper fresh ground black pepper cold water butter sugar egg yolk unsalted salted reduced fat creamer whole grain oats half & halves lowfat cheddar or Swiss cheeses light brown breadcrumbs apple juice frozen whipped topping ice cubes chocolate syrup optional garnish cooked carrots celery cucumbers green onions lemon wedges whipping whip heavy cream strawberry puree sherry wine fruit cocktail soda strawberries blueberries raspberries peaches pears bananas grapes watermelon seedless grapefruits pineapples orange slices mandarin oranges pineapple chunks cherries apples tart cherry pomegranate seeds extra virgin olive spread dried cranberries cinnamon crushed nuts coconut marshmallows angel food cake frosting cocoa dusted powdered sugar graham crackers swirl candy melts sea salt dark rum soy sauce maple whiskey lime zest fennel bulb dill weed shallot scallions yellow onion garlic cloves mint leaves basil chives parsley thyme oregano saffron turmeric coriander cumin cardamom ginger marjoram bay leaf peppercorns rose petals star anise licorice root tumeric goldenrod geranium tea clove rosemary elderberry spice pumpkin pie spices jalapeno peppers red bell peppers canned mushrooms chicken breast boneless skinless turkey wing tenderlo	Cottage Cheese Waffle Bites - Instructions per person: 1) Preheat oven to 400°F (around room temperature). Grease two baking sheets lightly using nonstick spray or butter . Melt some coconut oil over medium heat , add chopped onion & garlic if desired ; cook until softened slightly but not browned): Prepare batter by combining dry mixes [flour / oatmeals] separately then whisk together wet mixes[water + melted fat+ mixed vegetables]. Pour equal amounts into individual ramekins spread evenly apart onto greased sheet trays ensuring no overflowing occurs during boiling phase later b): Once bubbles appear around edges let rest about five minutes before adding more liquid gradually stirring well after every addition till smooth consistency achieved without lumps remaining c): Drop spoonful sized portions gently off counter surface directly upwards through small hole made earlier within parchment lined tray avoiding excessive air exposure between drops causing deflation affecting final product shape uniformity upon serving time ! Enjoy warm immediately post completion! Good luck crafting your delicious creation today :) Happy Cooking!!

Continued on next page

Table 9 – continued from previous page		
Model	**Original Output**	**With Allergen Substitution**
Phi-2	Cottage Cheese Waffle Recipe Ingredients : 1 cup all purpose white wheat flours 2 teaspoons kosher or coarse sea salt 3 large egg whites whole grain self raising dry instant yeast cheddar grated parmesan fresh ground black pepper caster sugar low fat plain yogurt buttermilk cold water butter sliced mushrooms olive leaves honey unsalted roasted walnuts maple syrup raspberry jam blueberries raspberries strawberry preserves lemon zest orange peel cinnamon stick ginger root dark rum double cream granulated sugar raw cane brown sugars coconut flakes nuts seeds dried apricot raisins breadcrumbs sunflower seed kernels crushed pecans powdered cocoa powder Instructions : preheat the oven to 180 degrees Celsius / 350 f / gas mark 4 & grease 6 nonstick mini muffin tins with cooking spray, such as PAM make this batter super quick you can combine everything into one bowl in order from easiest preparation step first - adding each ingredient at the end when it is ready so that they will blend well together without losing their individual flavours e.g. if your baking time depends on which appliance you use choose whichever has fastest cycle times! add any extra flavourings later too after mixing other things like fruit etc! don't worry about making mistakes though because there's always next week's	Cottage Cheese Waffle Breakfast Bowl Ingredients : gluten free pancake mix rice krispies frozen mixed berries almond butter powdered sugar vegan margarine nonfat ricotta fresh ground black pepper cinnamon maple syrup raw honey orange zest flax seed meal unsweetened shredded coconut granola chopped walnuts dried cranberries chia seeds cashews lemon juice lime peel paprika smoked sea salt cayenne cracked black peppercorns olive oil soy sauce white wine vinegar shallots fresh mint leaves extra virgin coconut cream low sodium soy broth apple cider vinegar dill weed capers onion flakes browning garlic powder jalapeno peppers corn tortillas avocado red bell pepper green chili scallion cucumber celery ribs tomato salad dressing mayonnaise parmesan dijon mustard french bread romaine lettuce baby carrots sliced strawberries blueberries whipped topping pomegranate molasses slivered almonds Instructions: preheat oven at 375 degrees fahrenheit or 190 Celsius degree for the breakfast bowl, take 1 tablespoon each raspberry jam & pineapple preserves & 2 tablespoons banana ice creams nectar|raspberry flavoring into your blender jar along with one cup whole nuts - crushed pine nut | macadamia's peanut | pistachio peanuts | hazelnut skins + / 3

E Domain-Specific Evaluation Metrics for Recipe Generation

Metric	Implementation Details
Ingredient Coverage	- Tokenize ingredients list - Match ingredients in instructions - Handle variations and forms - Calculate coverage ratio
Step Complexity	- Count distinct operations - Analyze step length and detail - Evaluate parameter specifications - Calculate complexity score
Recipe Coherence	- Build step dependency graph - Verify logical ordering - Check temporal consistency - Calculate coherence score
Temperature/Time	- Extract numerical values - Validate ranges per method - Check completeness - Calculate specification score

F Allergen Substitution Database for RAG-assisted System

Allergen Ingredient	Substitutes	Notes
Peanuts	Sunflower seed butter, almond butter, soy butter, pumpkin seed butter, cashew butter	Choose based on specific allergies. Similar protein content and texture.
Tree Nuts	Seeds, roasted chickpeas, coconut, pretzels, sunflower seeds	Ensure substitute is safe for specific nut allergy.
Milk	Oat milk, almond milk, soy milk, coconut milk, cashew milk	Oat milk works best for baking, coconut milk for curry dishes.
Eggs	Flax eggs, chia eggs, mashed banana, applesauce, commercial egg replacer	For binding: 1 egg = 1 tbsp ground flax + 3 tbsp water
Wheat	Almond flour, coconut flour, oat flour, rice flour, quinoa flour	May need to adjust liquid ratios when substituting.
Soy	Coconut aminos, chickpeas, hemp seeds, quinoa, pea protein	Coconut aminos work well for soy sauce replacement.
Fish	Hearts of palm, jackfruit, mushrooms, tempeh, seitan	Hearts of palm works great for fish-like texture.
Shellfish	King oyster mushrooms, hearts of palm, artichoke hearts, jackfruit, palm hearts	King oyster mushrooms provide similar texture to scallops.
Sesame	Poppy seeds, hemp seeds, flax seeds, sunflower seeds, pumpkin seeds	Similar nutty flavor profile.
Dairy	Coconut cream, cashew cream, nutritional yeast, vegan cheese, tahini	Nutritional yeast adds cheesy flavor.
Butter	Coconut oil, olive oil, avocado, applesauce, vegan butter	Coconut oil works best for baking.
Cheese	Nutritional yeast, cashew cheese, tofu, vegan cheese, hummus	Nutritional yeast adds umami flavor.
Cream	Coconut cream, cashew cream, silken tofu, oat cream, soy cream	Coconut cream works best for curry and soups.
Yogurt	Coconut yogurt, almond yogurt, soy yogurt, cashew yogurt, oat yogurt	Similar texture and tanginess.
Gluten	Xanthan gum, guar gum, psyllium husk, chia seeds, flax seeds	Important for binding in gluten-free baking.

G LLM-As-Judge Radar Charts for Prompt-based Method

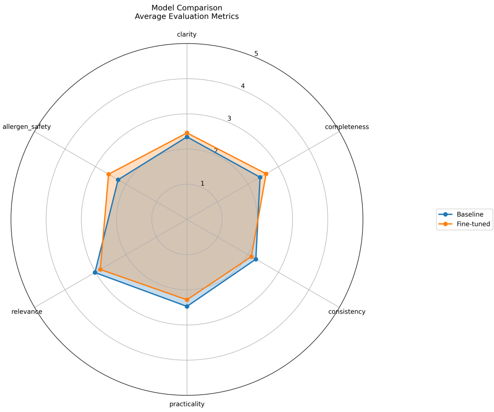

Figure 5: Comparison between Baseline and Fine-Tuned-SmolLm360

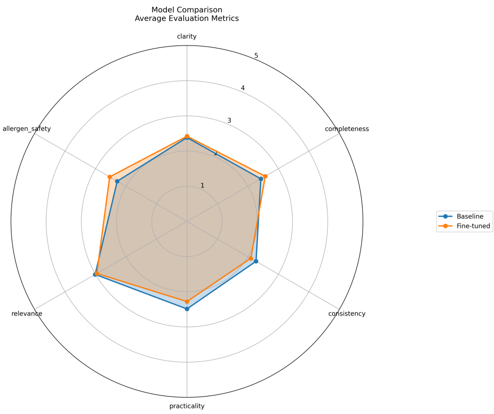

Figure 6: Comparison between Baseline and Fine-Tuned-SmolLm1.7B

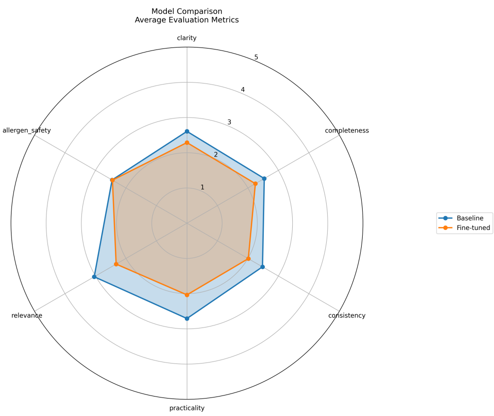

Figure 7: Comparison between Baseline and Fine-Tuned-Phi-2

H LLM-As-Judge Radar Charts for Rag-assisted Method

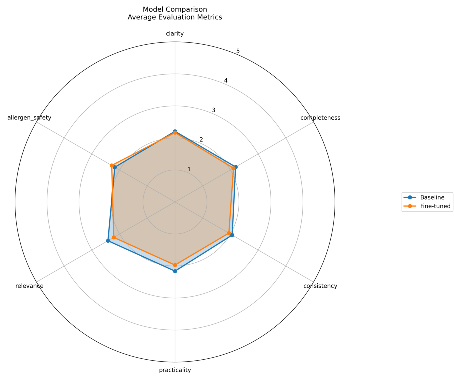

Figure 8: Comparison between Baseline and Fine-Tuned-SmolLm360

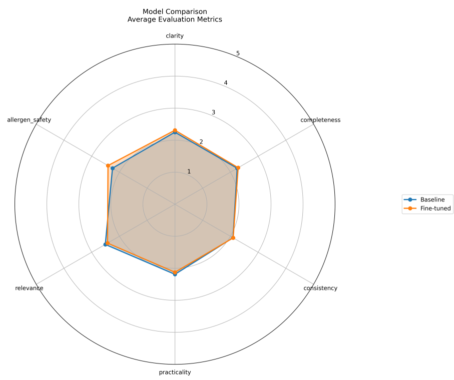

Figure 9: Comparison between Baseline and Fine-Tuned-SmolLm1.7B

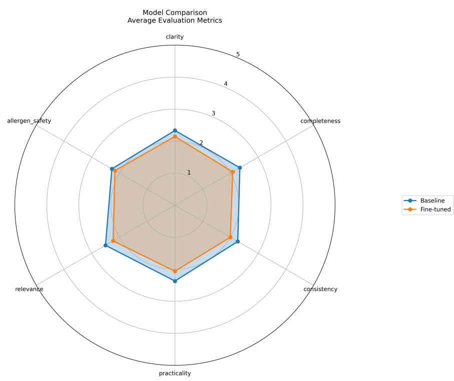

Figure 10: Comparison between Baseline and Fine-Tuned-Phi-2

I Training Configurations for Small-Scale Models

Parameter	Value
Batch Size	32
Learning Rate	2e-5, 1e-6 (GPT2)
Weight Decay	0.01
Warmup Steps	100
Gradient Accumulation	4
Mixed Precision*	fp16 or fp32
Optimizer	AdamW

Table 10: Training Configuration Details for Small-Scale Models

For Mixed Precision, we used both fp16 and fp32 due to dependency issues and limited computational resources.

J Training Configurations for Large-Scale Models

Parameter	Value
Batch Size	32
Learning Rate	2e-4
Weight Decay	0.01
Warmup Steps	100
Gradient Accumulation	4
Mixed Precision	fp16
Optimizer	`paged_adamw_8bit`

Table 11: Training Configuration Details for Large-Scale Models

K Hyper parameters for Generation in Prompt based Allergen Substitution

Parameter	Value
Max new tokens	256
Temperature	0.75
Top p	0.95
Do sample	True
No repeat ngram size	4
repetition penalty	1.3

Table 12: Hyper parameters for Generation in Prompt based Allergen Substitution

L Training Configurations for Large-Scale Models

Parameter	Value
Max new tokens	256
Temperature	0.75
Top p	0.95
Do sample	True
No repeat ngram size	4
repetition penalty	1.3

Table 13: Training Configuration Details for Large-Scale Models