Trip Weaver:
An Artificial Travel Agent on the Edge

Manolo Alvarez
Department of Electrical Engineering
Stanford University
manoloac@stanford.edu

Abstract

Travel is a fun, enriching experience that often succumbs to the burdensome weight of intricate planning and intractable financial estimates. This project attempts to relief that stress by creating an artificially intelligent agent capable of formulating comprehensive travel itineraries with cost approximations, tailored to individual preferences, and within the bounds of consumer hardware. The agent, a large language model (LLM), is tuned via Quantized Low-Rank Adapters (QLoRA), maximum-likelihood training on expert itineraries (SFT), and implicit reward models through Direct Preference Optimization (DPO). In evaluation, the 7 billion quantized parameter SFT and DPO models improve significantly over the baseline with the SFT model approximating the quantitative and qualitative performance of GPT-3.5.

1 Introduction

Travel planning, often a complex and time-consuming task, has the potential to evolve significantly with the advent of Artificial Intelligence (AI). The ability to generate detailed itineraries that cater to individual preferences, while managing budget constraints, has long been a desired feature of avid travelers like myself. In this paper I introduce an approach to fine-tune an open-sourced language model for such a task. More specifically, I demonstrate how to tune the 7-billion parameter Llama-2 language model via Quantized Low-Rank Adapters (QLoRA)[4], maximum-likelihood training on expert itineraries (SFT), and implicit reward models through Direct Preference Optimization (DPO)[3], to create personalized travel plans. The aim is to significantly reduce the planning burden for travelers, offering them tailor-made itineraries that align with their preferences and financial considerations, all within the practical constraints of consumer hardware.

This report delves into a method for dataset generation, model training and optimization, the challenges encountered, and the successes achieved. It offers a simple and effective approach to travel planning, making it more accessible, efficient, and enjoyable for travelers. The github repo can be found here: https://github.com/manolo-alvarez/TripWeaver

2 Related Work

There are a few websites that offer AI generated travel itineraries, most of which plug into some version of GPT, but no work has been done to fine-tune open-sourced models for this task.

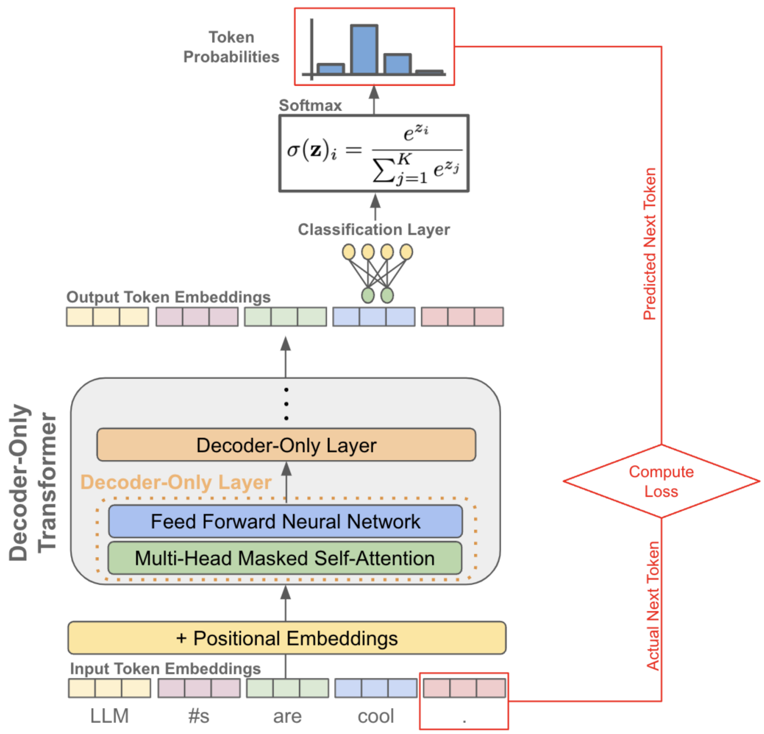


Figure 1: LLama-2 Transformer Architecture

Constitutional AI [5] uses a self-improvement method for training AI with a set of predefined rules or principles without human-generated labels. The training process consists of two phases: a supervised learning (SL) phase, where the AI generates self-critiques and revisions based on an initial model, and a reinforcement learning (RL) phase, where the AI's choices are evaluated and used to train a preference model that serves as a reward signal. In this project, I use a rubric analogous to the constitution, but instead of 'RL from AI Feedback' (RLAIF), I explore DPO from AI Scoring without the critiquing steps.

3 Method

This project builds on top of the 7 billion parameter model from the open-sourced Llama-2 suite [2] by Meta, and directly leverages the work presented by the authors of the LoRA [1], QLoRA [4], and DPO [3] papers. Next, we'll analyze these techniques independently and then dig into the full method.

3.1 Llama-2

Llama-2's model architecture follows the standard transformer architecture illustrated in Figure 1

3.2 LoRA

In LoRA 2, the weights of the neural network are modified by adding a low-rank matrix. If we consider a weight matrix W in the attention layers of Llama-2, the LoRA modification is applied as $W' = W + BA$, where W' is the modified weight matrix. B and A are the low-rank matrices with a rank factor of r. This means if W is of dimension $d \times k$, then B is of dimension $d \times r$ and A is of dimension $r \times k$.

During training, only the parameters of the low-rank matrices B and A are updated. The original parameters of Llama-2 remain frozen. To apply the LoRA modification, we add the BA product to the original weight matrices of the self-attention layers Q, K, and V.

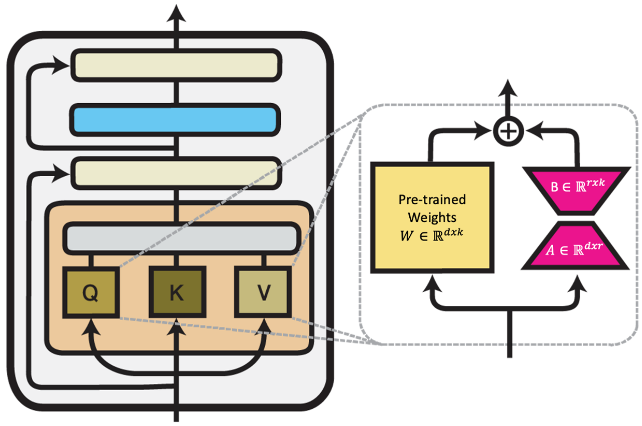


Figure 2: Low-Rank Adapters

QLoRA [4], 4-bit quantizes the model's weights on top of LoRA 2, shrinking the model size by a factor of 4 at the expense of parameter precision. In practice, quantization can slightly hinder model performance, but it was a conscious trade-off I made in order to fit the models I was exploring within personal hardware.

3.3 DPO

The traditional approach in optimizing human preferences through reinforcement learning (RL) involves using an auxiliary reward model to fine-tune the main model, encouraging it to produce more high-reward outputs and fewer low-reward ones. The Direct Policy Optimization (DPO) method streamlines this process by directly optimizing the language model using preference data. It achieves this through a novel analytical approach that converts the RL loss involving both reward and reference models into a loss solely based on the reference model, eliminating the need for a complex RL-based optimization process.

To understand DPO, it is helpful to analyze the gradient of the loss function with respect to the parameters θ:

$$\nabla_\theta L_{DPO}(\pi_\theta; \pi_{ref}) =$$

$$-\beta \mathbb{E}_{(x,y_w,y_l)\sim D}\left[\underbrace{\sigma(\hat{r}_\theta(x,y_l) - \hat{r}_\theta(x,y_w))}_{\textit{higher when reward estimate is wrong}}\left[\underbrace{\nabla_\theta \log \pi(y_w|x)}_{\textit{increase likelihood of } y_w} - \underbrace{\nabla_\theta \log \pi(y_l|x)}_{\textit{decrease likelihood of } y_l}\right]\right] \tag{1}$$

where $\hat{r}_\theta(x,y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}$ is the reward implicitly defined by the language model π_θ and reference model $\pi_{ref}(y|x)$. As the labels in equation 1, note, the gradient of the loss function will increase the likelihood of the preferred completions, decrease the likelihood of the dis-preferred completions, and scale the update by how much higher the implicit reward model $\hat{r}_\theta$ rates the dis-preferred completions [3].

3.4 Trip Weaver

With limited time and resources, the options for generating a dataset from scratch and training a model to match human preferences were few. Most people that embark on a long trip will generate some kind of itinerary. Unfortunately, most of those itineraries remain private and miss significant detail. For those reasons, I opted for the most time-efficient method of data collection that I believed could match the quality of a human-generated dataset with labels.

In the method below, I describe a process for scoring itineraries based off a rubric. While I could have had GPT choose the winning and losing itineraries directly, I opted for it to score these instead, thinking that it would be an easier "reasoning" task for the language model over the alternative.

Scoring also made it easier to rank multiple itineraries generated by a single prompt, if I chose to do so later on.

Let's walk through the method, as depicted in figure 3, step-by-step:

1. **Generate** 1.6k pairs of GPT-3.5-Turbo-1106 (GPT) completions given a standard prompt with variables: city, budget, and duration of trip:

   "I will be traveling to {city} from {start date} through {end date}. Generate an itinerary that details activities, transportation between destinations, approximate expenses per activity, breakfast location, lunch location, dinner location, is flexible, and doesn't cost more than {total budget} total (not including flights and accommodation)."

   where the city is sampled from the latest Wikipedia list of the 100 most visited places in the world, the start date is a random date between January 1st, 2024 and December 31st, 2025, the end date is a random date 2-5 days after the start date, and total budget is a random number between $500 and $1500 USD.

2. **Score** each itinerary with GPT given the standard rubric:

   - Activities and Attractions (35 points) - Variety of activities, including cultural, recreational, and historical experiences. Inclusion of must-see attractions. Balance between structured and free time.
   - Transportation (15 points) - Well-planned transportation between destinations. Efficiency and convenience of travel options.
   - Budget and Expenses (25 points) - Clear cost breakdown and financial planning. Appropriateness of expenses based on the traveler's budget. Strategies for cost-saving, if applicable.
   - Local Cuisine and Dining (10 points) - Inclusion of authentic local dining experiences. Recommendations for must-try dishes.
   - Cultural Immersion (10 points) - Opportunities for interacting with locals. Exploration of local customs and traditions. Integration of cultural experiences into the itinerary.
   - Flexibility (5 points) - Ability to adapt the itinerary based on unexpected circumstances. Contingency plans for weather-related or unforeseen issues.

   Total Score: 100 points

3. **Compare** the scores of the itineraries in each pair. The one with the higher score is the "winner" and the other is the "loser".

4. **Train** the 7 billion parameter Llama-2 language model (base model) with QLoRA to maximize the likelihood of generating the winning itinerary given its corresponding prompt. This is the Supervised Fine-tuning (SFT) step.

5. **Train** the base model with QLoRA, starting with the trained adapters from the SFT step, via DPO. The objective is to maximize the likelihood of the model generating the winning itinerary and decrease the likelihood of it generating the loosing itinerary, given its corresponding prompt.

4 Evaluation

The "baseline" model for this project was the llama-2-7b quantized model. This model served as the initial comparison point for subsequent enhancements and evaluations in the Supervised Fine-tuning (SFT) and Direct Preference Optimization (DPO) steps.

4.1 Metrics

The primary metrics used to evaluate model performance were the average and standard error of scores across five random itineraries on the rubric graded by the GPT model 4, and my subjective evaluations. Given how narrow the task here is, I did not believe other common metrics used to evaluate LLM performance would provide any meaningful insight.

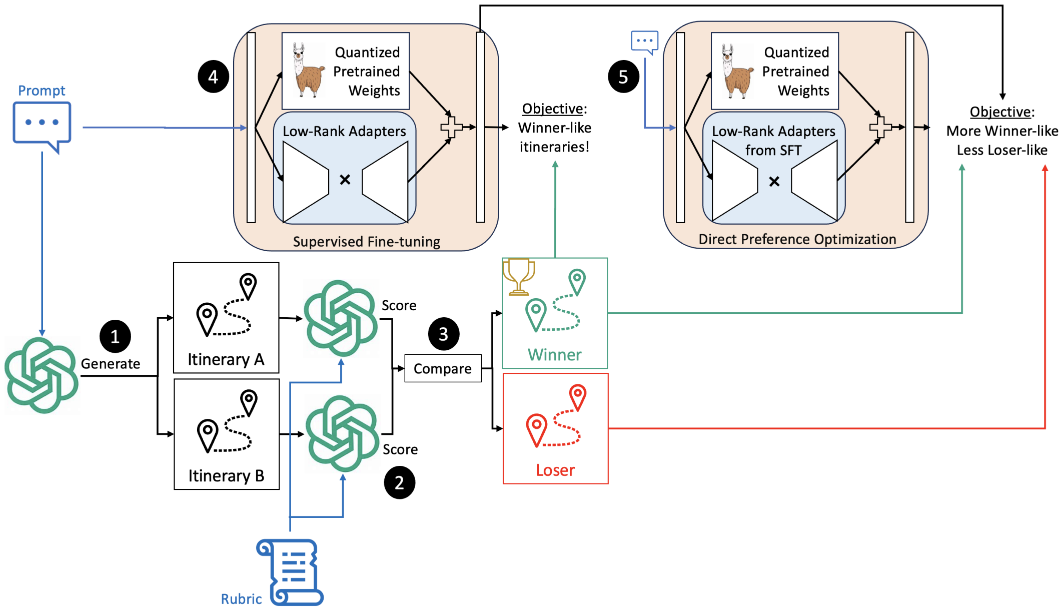


Figure 3: Method for Data Collection & Training

	Baseline	SFT	DPO	GPT-3.5
Average	0.0	91.8	81.4	94.0
Standard Error	0.0	2.2	4.9	1.0

Figure 4: Average & standard error of itinerary scores across models

4.2 Baseline Performance

When prompted with the same standard prompt described in earlier sections, the baseline model struggled to generate coherent itineraries. In the example completion in figure 5, the model was prompted to generate an itinerary for Austin, TX and it ended up talking about New York. The itineraries were so bad that GPT misunderstood the task when asked for a score. It ended up creating a new itinerary and scoring that one instead. For that reason, I felt it fair to score it a 0 myself, as seen on figure 4.

The completions did improve somewhat on a slightly modified prompt but for consistent evaluation across the models, I decided to show the results on the standard prompt.

4.3 SFT Performance

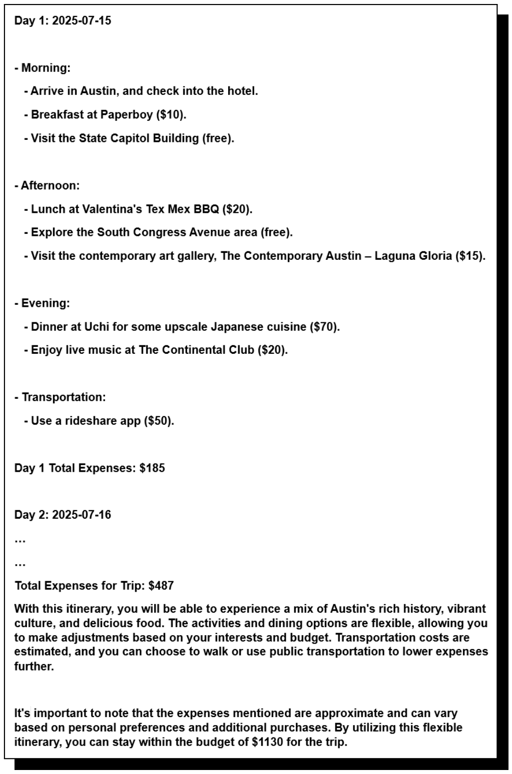
After just 100 epochs, the SFT model demonstrated significant qualitative improvements over the baseline, aligning closely in performance with GPT-3.5-Turbo-1106. Quantitatively, the model scored an average of 91.8 with a 2.2 standard error out of the 100 points possible on the rubric 4.

4.4 DPO Performance

Unexpectedly, the DPO model notably worsened from the SFT model in both quality and quantitative performance. While it did not degenerate back to the baseline, it is clear by the sampled itineraries that it did not align well with the intended objective.

4.5 Analysis

The base model was useless for this task. Often spewing code or incoherent text. So it was nice to observe remarkable improvement on quantitative and qualitative performance after a brief bout of supervised fine-tuning (SFT). While the itineraries lack a bit of detail around activities and may



Day 1: 2025-07-15

- Morning:
  - Arrive in Austin, and check into the hotel.
  - Breakfast at Paperboy ($10).
  - Visit the State Capitol Building (free).

- Afternoon:
  - Lunch at Valentina's Tex Mex BBQ ($20).
  - Explore the South Congress Avenue area (free).
  - Visit the contemporary art gallery, The Contemporary Austin – Laguna Gloria ($15).

- Evening:
  - Dinner at Uchi for some upscale Japanese cuisine ($70).
  - Enjoy live music at The Continental Club ($20).

- Transportation:
  - Use a rideshare app ($50).

Day 1 Total Expenses: $185

Day 2: 2025-07-16

…

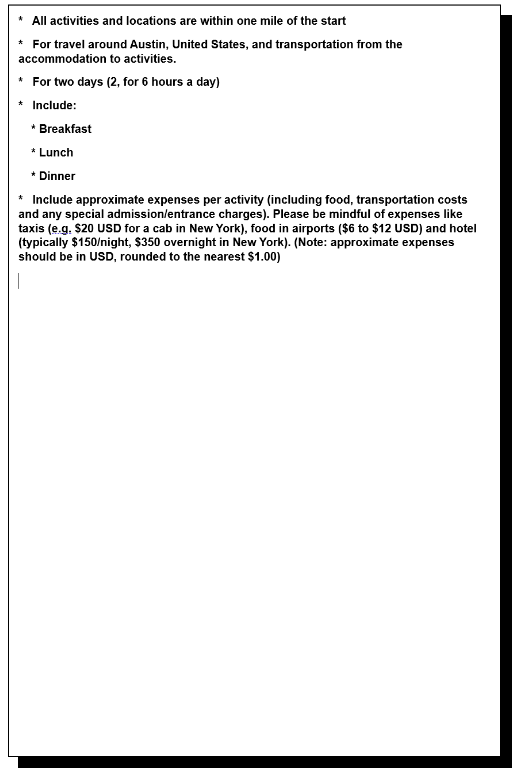
…

Total Expenses for Trip: $487

With this itinerary, you will be able to experience a mix of Austin's rich history, vibrant culture, and delicious food. The activities and dining options are flexible, allowing you to make adjustments based on your interests and budget. Transportation costs are estimated, and you can choose to walk or use public transportation to lower expenses further.

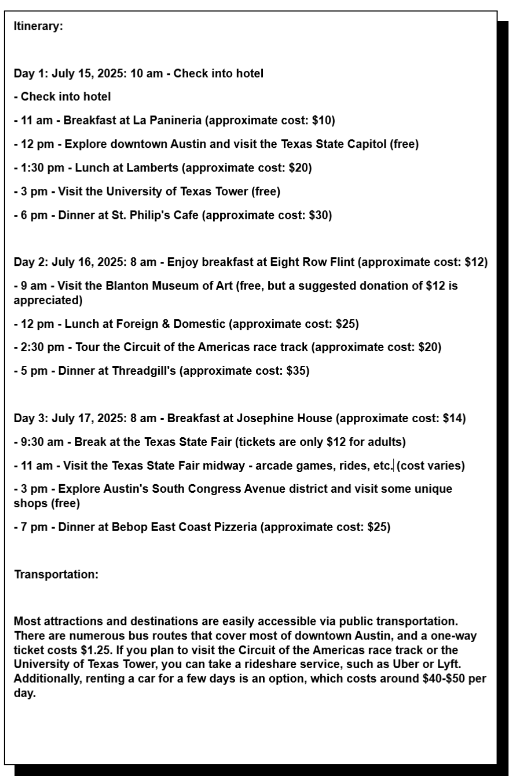
It's important to note that the expenses mentioned are approximate and can vary based on personal preferences and additional purchases. By utilizing this flexible itinerary, you can stay within the budget of $1130 for the trip.

(a) GPT-3.5-Turbo-1106



* All activities and locations are within one mile of the start
* For travel around Austin, United States, and transportation from the accommodation to activities.
* For two days (2, for 6 hours a day)
* Include:
  * Breakfast
  * Lunch
  * Dinner
* Include approximate expenses per activity (including food, transportation costs and any special admission/entrance charges). Please be mindful of expenses like taxis (e.g. $20 USD for a cab in New York), food in airports ($6 to $12 USD) and hotel (typically $150/night, $350 overnight in New York). (Note: approximate expenses should be in USD, rounded to the nearest $1.00)

(b) Baseline



Itinerary:

Day 1: July 15, 2025: 10 am - Check into hotel

- Check into hotel
- 11 am - Breakfast at La Panineria (approximate cost: $10)
- 12 pm - Explore downtown Austin and visit the Texas State Capitol (free)
- 1:30 pm - Lunch at Lamberts (approximate cost: $20)
- 3 pm - Visit the University of Texas Tower (free)
- 6 pm - Dinner at St. Philip's Cafe (approximate cost: $30)

Day 2: July 16, 2025: 8 am - Enjoy breakfast at Eight Row Flint (approximate cost: $12)

- 9 am - Visit the Blanton Museum of Art (free, but a suggested donation of $12 is appreciated)
- 12 pm - Lunch at Foreign & Domestic (approximate cost: $25)
- 2:30 pm - Tour the Circuit of the Americas race track (approximate cost: $20)
- 5 pm - Dinner at Threadgill's (approximate cost: $35)

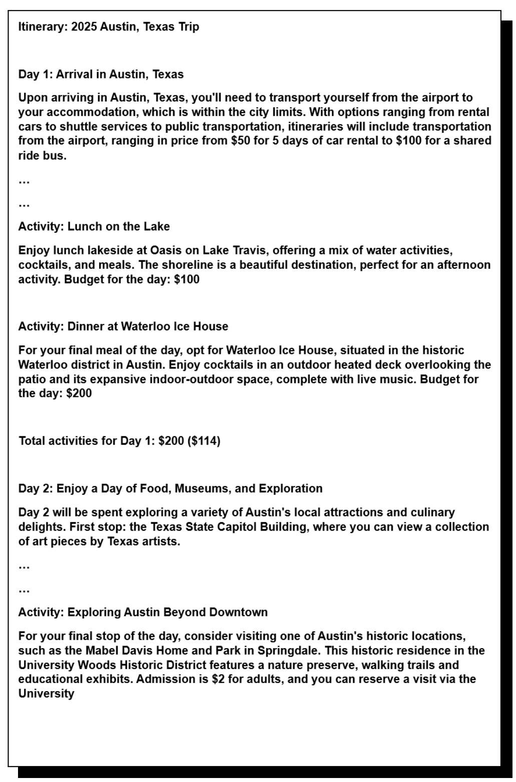
Day 3: July 17, 2025: 8 am - Breakfast at Josephine House (approximate cost: $14)

- 9:30 am - Break at the Texas State Fair (tickets are only $12 for adults)
- 11 am - Visit the Texas State Fair midway - arcade games, rides, etc. (cost varies)
- 3 pm - Explore Austin's South Congress Avenue district and visit some unique shops (free)
- 7 pm - Dinner at Bebop East Coast Pizzeria (approximate cost: $25)

Transportation:

Most attractions and destinations are easily accessible via public transportation. There are numerous bus routes that cover most of downtown Austin, and a one-way ticket costs $1.25. If you plan to visit the Circuit of the Americas race track or the University of Texas Tower, you can take a rideshare service, such as Uber or Lyft. Additionally, renting a car for a few days is an option, which costs around $40-$50 per day.

(c) SFT



Itinerary: 2025 Austin, Texas Trip

Day 1: Arrival in Austin, Texas

Upon arriving in Austin, Texas, you'll need to transport yourself from the airport to your accommodation, which is within the city limits. With options ranging from rental cars to shuttle services to public transportation, itineraries will include transportation from the airport, ranging in price from $50 for 5 days of car rental to $100 for a shared ride bus.

…

…

Activity: Lunch on the Lake

Enjoy lunch lakeside at Oasis on Lake Travis, offering a mix of water activities, cocktails, and meals. The shoreline is a beautiful destination, perfect for an afternoon activity. Budget for the day: $100

Activity: Dinner at Waterloo Ice House

For your final meal of the day, opt for Waterloo Ice House, situated in the historic Waterloo district in Austin. Enjoy cocktails in an outdoor heated deck overlooking the patio and its expansive indoor-outdoor space, complete with live music. Budget for the day: $200

Total activities for Day 1: $200 ($114)

Day 2: Enjoy a Day of Food, Museums, and Exploration

Day 2 will be spent exploring a variety of Austin's local attractions and culinary delights. First stop: the Texas State Capitol Building, where you can view a collection of art pieces by Texas artists.

…

…

Activity: Exploring Austin Beyond Downtown

For your final stop of the day, consider visiting one of Austin's historic locations, such as the Mabel Davis Home and Park in Springdale. This historic residence in the University Woods Historic District features a nature preserve, walking trails and educational exhibits. Admission is $2 for adults, and you can reserve a visit via the University

(d) DPO

Figure 5: Example Completions (Shortened for Visualization)

have minor errors here and there, their structure are on par and sometimes superior to that of GPT. Creating a nice free alternative to GPT and other online plug-ins to it.

Now, clearly, the DPO model performance did not align with my initial expectations, and I believe the reason is rooted in the dataset. The itineraries in the dataset were generated by GPT. They are all of decent quality. However, upon inspection, I noticed that the model was either too generous on or flawed in the reasoning behind the scores. Resulting in the distribution of scores across the dataset to be narrowly clustered around 75 and 95, and unaligned with what I, a human, would've scored them. Now, it could be that I was not specific enough in the rubric or that I did not allow the model to engage in a long enough chain of thought to reason appropriately. Either way, the scoring method is effectively the implicit reward function in DPO. If the method is flawed, it is likely that the dataset has chosen and rejected labels on itineraries that are unaligned with true human preferences. A weak evidence of this, is that when I narrowed down the dataset to itinerary pairs that had a difference in scores greater than 5, the model improved.

In the end, these are the results of a very time and resource-limited effort. For that reason, I believe these are great results with a large potential for improvement.

5 Conclusion

This project demonstrated the feasibility and potential of using an open-sourced language model (LM) to generate travel itineraries with consumer hardware. Employing Quantized Low-Rank Adapters (QLoRA), training for maximum likelihood (SFT) on expert itineraries, and training with implicit reward models through Direct Preference Optimization (DPO) resulted in models that aligned more closely with human preferences in travel planning. The key findings include:

1. **Effectiveness of Supervised Fine-Tuning (SFT)**: A short bout of SFT can significantly improve a foundational model's ability to generate coherent and practical itineraries. This improvement was evident both qualitatively and quantitatively, indicating the model's enhanced capacity to understand and execute the specific task of travel itinerary generation.
2. **Challenges with Direct Preference Optimization (DPO)** Contrary to expectations, training with DPO did not yield better results over the SFT model. This was likely not attributed to the DPO method itself but to flaws in the scoring method used to compare and align itineraries with human-like preferences.
3. **Importance of Dataset Quality** The project highlighted the critical role of dataset quality in training AI models, especially for tasks that require a nuanced understanding of human preferences. The slight improvement observed when refining the dataset suggests that more precise scoring and a broader range of quality in the itineraries could enhance the model's performance in future iterations.
4. **Potential for Further Improvement** Despite the limitations in time and resources, the project's outcomes were promising, showcasing the potential of AI in automating complex tasks like travel planning. Future work with more refined datasets and context-rich prompts could yield even better results.

In conclusion, this project underscores the significant potential of AI in the realm of personalized travel planning. While there are challenges to overcome, particularly in dataset generation, the progress made in this project serves as a foundation for future advancements in this domain.

References

[1] Phillip Wallis Zeyuan Allen-Zhu Yuanzhi Li Shean Wang Lu Wang Weizhu Chen Edward J. Hu, Yelong Shen. Lora: Low-rank adaptation of large language models, 2021.

[2] Kevin Stone Peter Albert Amjad Almahairi Yasmine Babaei Nikolay Bashlykov Soumya Batra Prajjwal Bhargava Shruti Bhosale Dan Bikel Lukas Blecher Cristian Canton Ferrer Moya Chen Guillem Cucurull David Esiobu Jude Fernandes Jeremy Fu Wenyin Fu Brian Fuller Cynthia Gao Vedanuj Goswami Naman Goyal Anthony Hartshorn Saghar Hosseini Rui Hou Hakan Inan Marcin Kardas Viktor Kerkez Madian Khabsa Isabel Kloumann Artem Korenev Punit Singh Koura Marie-Anne Lachaux Thibaut Lavril Jenya Lee Diana Liskovich Yinghai Lu Yuning Mao Xavier Martinet Todor Mihaylov Pushkar Mishra Igor Molybog Yixin Nie Andrew Poulton Jeremy Reizenstein Rashi Rungta Kalyan Saladi Alan Schelten Ruan Silva Eric Michael Smith Ranjan Subramanian Xiaoqing Ellen Tan Binh Tang Ross Taylor Adina Williams Jian Xiang Kuan Puxin Xu Zheng Yan Iliyan Zarov Yuchen Zhang Angela Fan Melanie Kambadur Sharan Narang Aurelien Rodriguez Robert Stojnic Sergey Edunov Thomas Scialom Hugo Touvron, Louis Martin. Llama 2: Open foundation and fine-tuned chat models, 2023.

[3] Eric Mitchell Stefano Ermon Christopher D. Manning Chelsea Finn Rafael Rafailov, Archit Sharma. Direct preference optimization: Your language model is secretly a reward model, 2023.

[4] Ari Holtzman Luke Zettlemoyer Tim Dettmers, Artidoro Pagnoni. Qlora: Efficient finetuning of quantized llms, 2023.

[5] Sandipan Kundu Amanda Askell Jackson Kernion Andy Jones Anna Chen Anna Goldie Azalia Mirhoseini Cameron McKinnon Carol Chen Catherine Olsson Christopher Olah Danny Hernandez Dawn Drain Deep Ganguli Dustin Li Eli Tran-Johnson Ethan Perez Jamie Kerr Jared Mueller Jeffrey Ladish Joshua Landau Kamal Ndousse Kamile Lukosuite Liane Lovitt Michael Sellitto Nelson Elhage Nicholas Schiefer Noemi Mercado Nova DasSarma Robert Lasenby Robin Larson Sam Ringer-Scott Johnston Shauna Kravec Sheer El Showk Stanislav Fort Tamera Lanham Timothy Telleen-Lawton Tom Conerly Tom Henighan Tristan Hume Samuel R. Bowman Zac Hatfield-Dodds Ben Mann Dario Amodei Nicholas Joseph Sam McCandlish Tom Brown Jared Kaplan Yuntao Bai, Saurav Kadavath. Constitutional ai: Harmlessness from ai feedback, 2022.