Residual Activation Steering

Ryan, Julian, Alex, Mame Coumba, Bill

Inspiration

Refusal in LLMs is mediated by a single direction

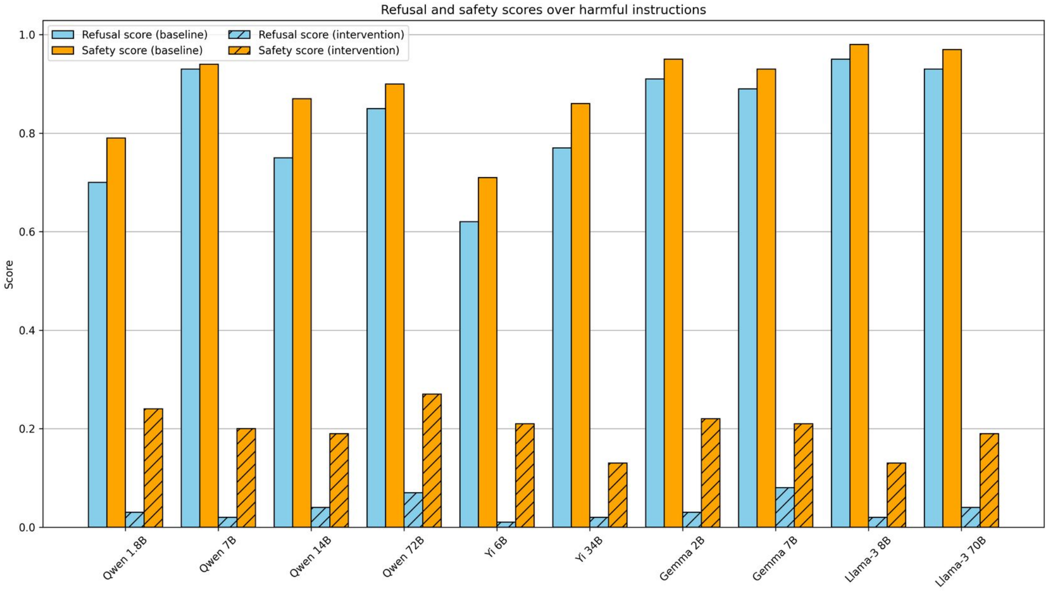
An inference-time intervention “which effectively jailbreaks the model without requiring any fine-tuning”

Refusal in LLMs is mediated by a single direction

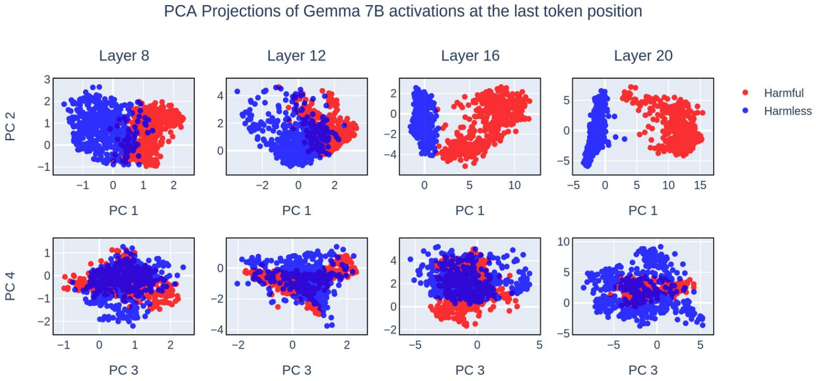
Steering overview

Removing refusal direction

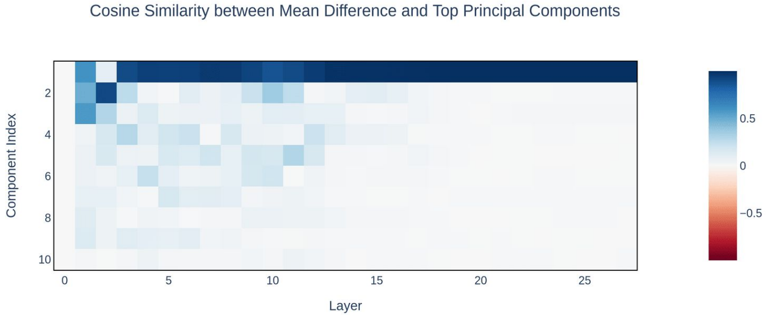
$$c'_{\text{out}} \leftarrow c_{\text{out}} - (c_{\text{out}} \cdot \hat{r})\hat{r}$$



Refusal and safety scores over harmful instructions



PCA Projections of Gemma 7B activations at the last token position



Cosine Similarity between Mean Difference and Top Principal Components

Task

Can we generalize this method to ablate arbitrary concepts from a pre-trained model?

Example: High School Economics

Motivation

Goal: Make a large language model incapable of answering specific topic questions or performing a specific task, while maintaining high performance in other areas

Use Cases:

Ablate potentially dangerous concepts like bomb-making or cybersecurity offense

Explore how models represent specific concepts internally

Custom Dataset

Generated By GPT-4

Curated manually to avoid hallucination, repetition

75 High-School Level Macroeconomics Questions

- Explain how an increase in consumer confidence impacts aggregate demand and overall economic performance.
- Describe the relationship between unemployment and inflation using the Phillips curve.
- What are the primary tools the Federal Reserve uses to implement monetary policy?

75 Generic Requests

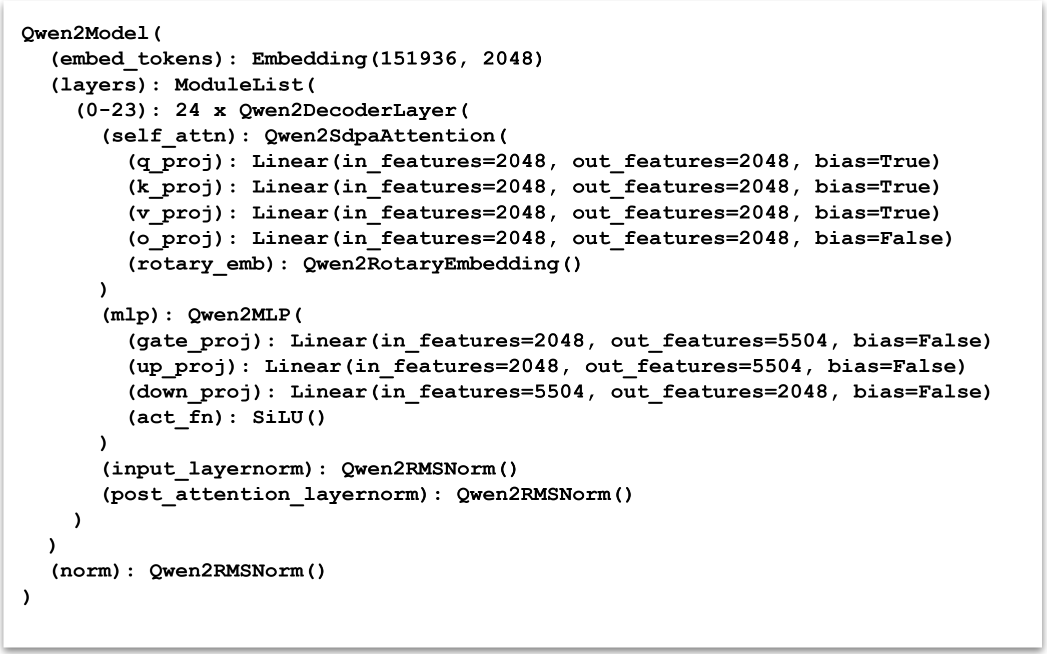
- Can you suggest some interesting novels to read for summer vacation?
- What's the best way to organize my study schedule for final exams?
- Could you explain the basics of photography composition?

Model: Qwen-1.5-1.8B-Chat

Qwen

- Decoder-only Transformer
- Open-source
- High Performance
- Flexible small model sizes for testing

Alibaba Cloud



```
Qwen2Model(
  (embed_tokens): Embedding(151936, 2048)
  (layers): ModuleList(
    (0-23): 24 x Qwen2DecoderLayer(
      (self_attn): Qwen2SdpaAttention(
        (q_proj): Linear(in_features=2048, out_features=2048, bias=True)
        (k_proj): Linear(in_features=2048, out_features=2048, bias=True)
        (v_proj): Linear(in_features=2048, out_features=2048, bias=True)
        (o_proj): Linear(in_features=2048, out_features=2048, bias=False)
        (rotary_emb): Qwen2RotaryEmbedding()
      )
      (mlp): Qwen2MLP(
        (gate_proj): Linear(in_features=2048, out_features=5504, bias=False)
        (up_proj): Linear(in_features=2048, out_features=5504, bias=False)
        (down_proj): Linear(in_features=5504, out_features=2048, bias=False)
        (act_fn): SiLU()
      )
      (input_layernorm): Qwen2RMSNorm()
      (post_attention_layernorm): Qwen2RMSNorm()
    )
  )
  (norm): Qwen2RMSNorm()
)
```

Methodology

- **Cache activations**
- **Compute topic direction**
- **Ablate model at all layers**
    - Implemented using PyTorch hooks

1. Cache activations from inference with **dataset of example phrases**
2. Compute steering vector (or “topic direction”) by the difference between **general queries** and **topic queries**
3. Select **specific layers** to ablate the model
4. **Perform inference/forward** pass with the given data
5. Calculate relevant BLEU and ROGUE metrics

Generic Prompts

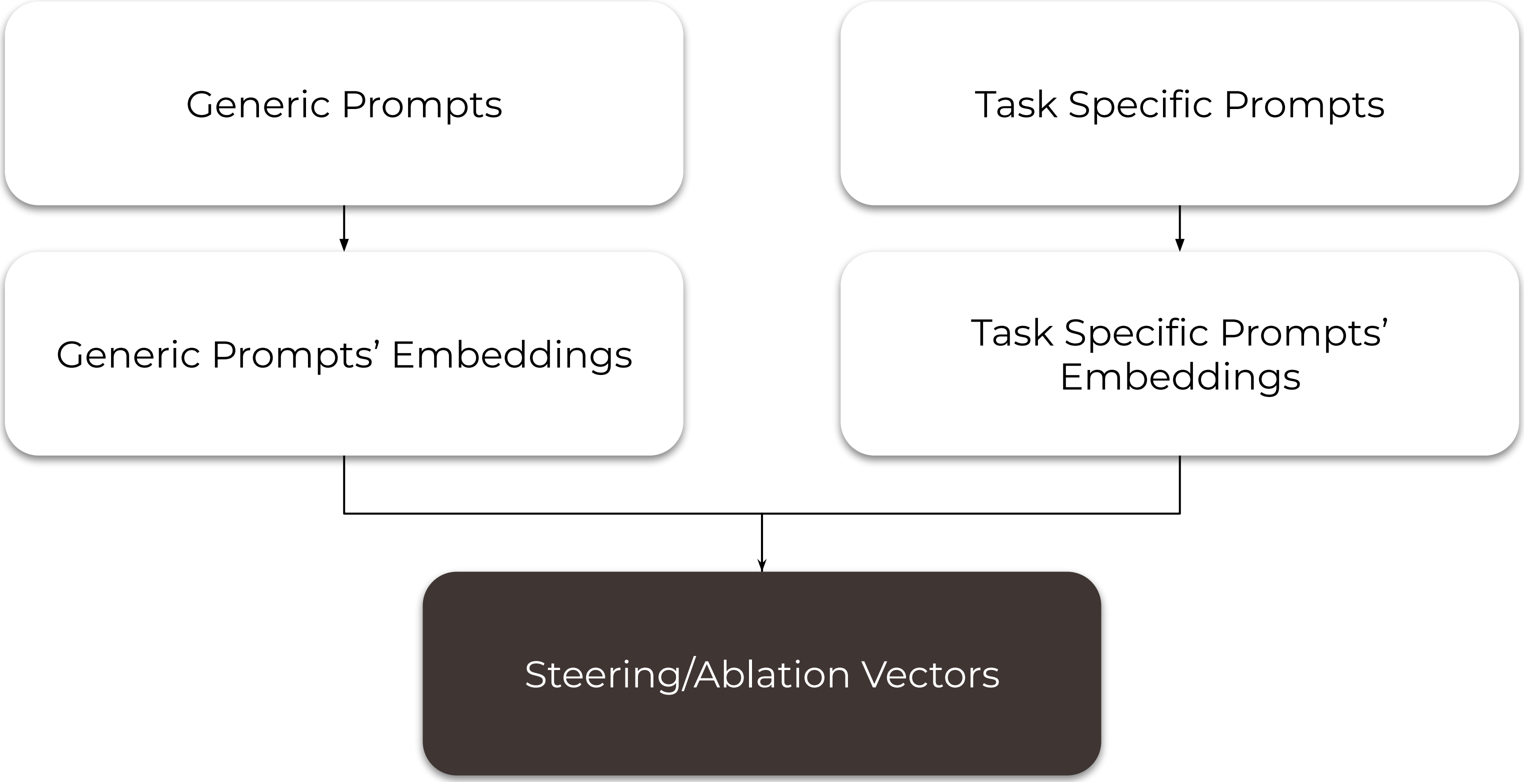
Task Specific Prompts

Generic Prompts

Generic Prompts’ Embeddings

Task Specific Prompts

Task Specific Prompts’ Embeddings



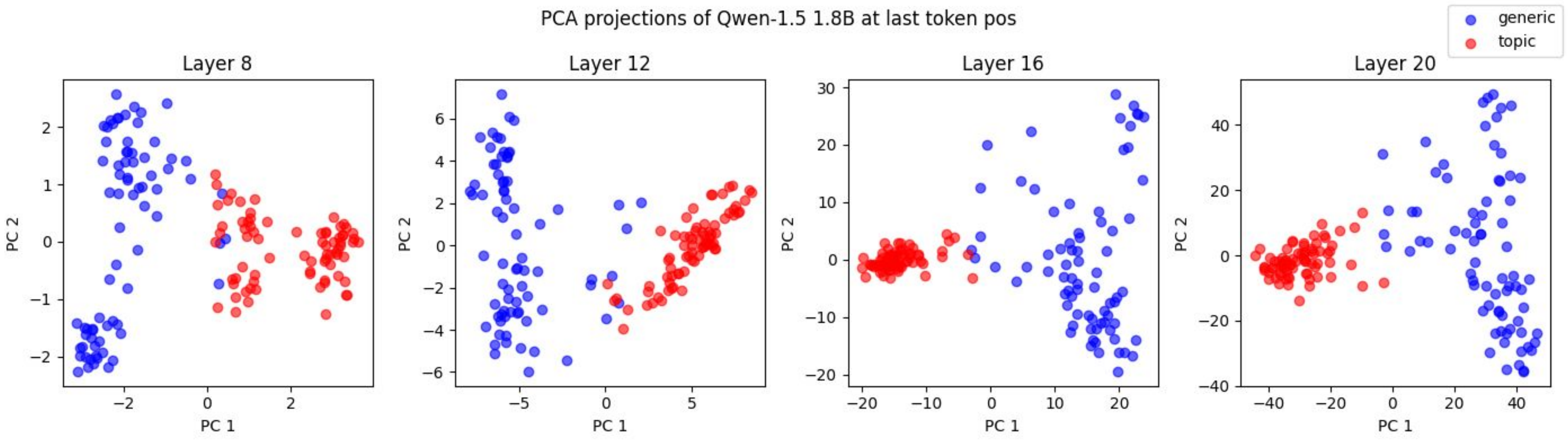
Generic Prompts

Task Specific Prompts

Generic Prompts’ Embeddings

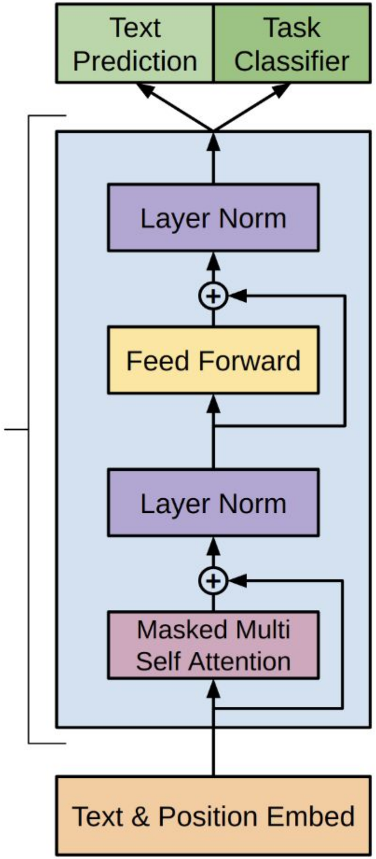
Task Specific Prompts’ Embeddings

Steering/Ablation Vectors



Visualizing residual stream activations for economic and generic prompts

2048-dim projected to 2D using PCA

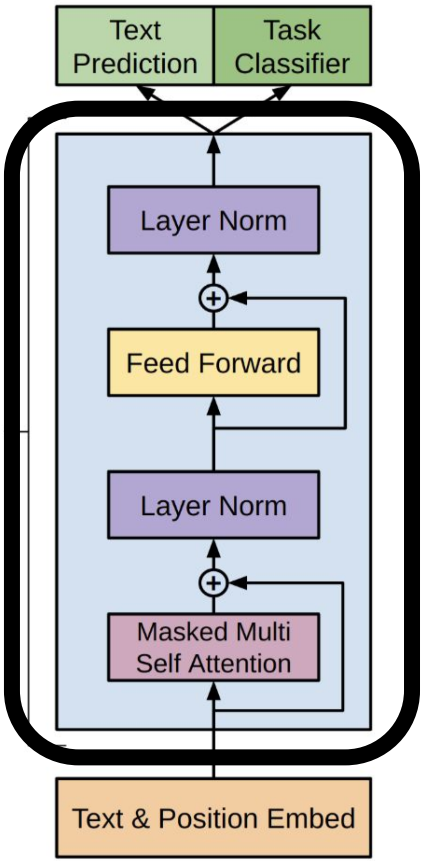


Implementation: PyTorch Hooks

```
hook(module, input, output) -> None or modified output
```

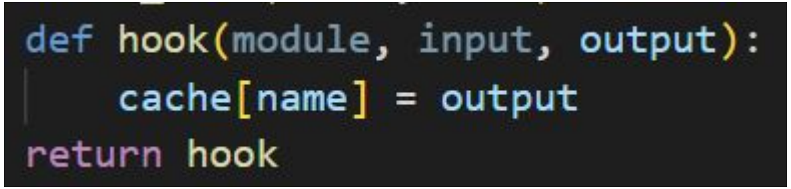
Implementation: PyTorch Hooks

`hook(module, input, output) -> None or modified output`

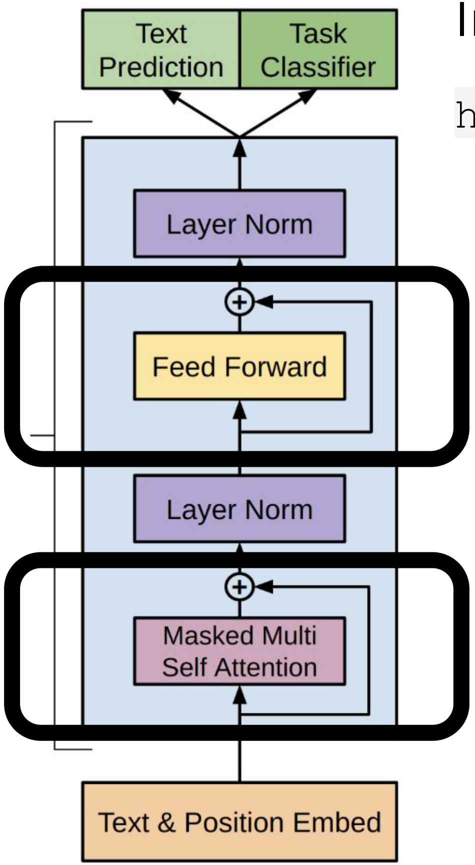


Caching activations

Hook each transformer layer
Save activations to cache



```
def hook(module, input, output):
    cache[name] = output
return hook
```



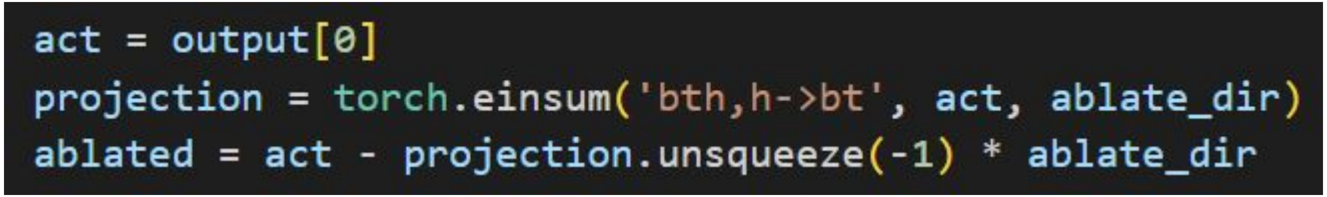
Implementation: PyTorch Hooks

```
hook(module, input, output) -> None or modified output
```

Ablating direction

Hook MLP and Self-attention modules

1. Project onto ablation direction
2. Subtract projection from activation



```
act = output[0]
projection = torch.einsum('bth,h->bt', act, ablate_dir)
ablated = act - projection.unsqueeze(-1) * ablate_dir
```

Results

- **Qualitative**
- **Quantitative**
  - Multiple choice answering
  - Automated generation metrics
  - GPT-4 scoring

User: What are the primary tools the Federal Reserve uses to implement monetary policy?

The Federal Reserve is responsible for implementing monetary policy, which is the management of the nation's money supply and economic activity through a series of decisions made at its meetings. Some of the primary tools the Federal Reserve uses to implement monetary policy include:

1. Open Market Operations (OMO): The Fed uses OMO to buy or sell U.S. government bonds, including Treasury bills, notes, and mortgage-backed securities, in an effort to influence short-term interest rates. This tool aims to control inflation by adjusting the amount of cash that the central bank purchases or sells ...

Default Qwen-1.8B Chat Output

The Federal Reserve is responsible for managing several important tasks when it comes to implementing great deals of money management. Some of the key tools that they use include:

1. Setting up an automated savings plan: One of the best ways to save money and keep it on track is by setting up an automatic savings plan. You can set aside a certain amount of money each month or year, depending on your budget and financial goals, and have those funds automatically transferred to a savings account. This way, you'll be saving consistently without even thinking about it!

2. Creating a budget: Creating a budget is essential for anyone who wants to manage their finances ...

With Ablation: Layers [6, 16, 22]

User: Describe the relationship between unemployment and inflation using the Phillips curve.

The Phillips curve is a graphical representation of the relationship between unemployment and inflation in the United States during different periods of time. It shows that as unemployment decreases, inflation also tends to decrease, but there is a point at which unemployment becomes too low (below the natural rate) that inflation starts to rise.

At the base of the curve, unemployment is below the natural rate, and as a result, inflation remains low or even negative. This occurs because businesses have sufficient capital to invest and expand without relying on government spending for monetary policy, so they are more likely to increase wages and productivity through technological advancements or other means, leading to lower prices for ...

Default Qwen-1.8B Chat Output

The perfect combination of getting hired without looking for work (preventing job searching) with finding great paying jobs that will help you keep up with living expenses after getting hired (reducing unnecessary expenses and saving money) can be very beneficial to managing your finances while trying to find employment and reducing your stress about not being able to get hired. Here's an approach to creating a plan that can help you save money, reduce your stress, and maintain good financial health:

1. Start by updating your resume: Make sure it's tailored to your desired career path, highlighting your skills, achievements, and any relevant experiences. A well-organized and ...

With Ablation: Layers [6, 16, 22]

This is working! (at least somewhat)

Ablation clearly harms model's ability to explain economic topics

- Instead, focuses on related ideas like personal finance advice

This matches our desired goal

So how can we measure:

- Effectiveness of ablation
- Model retaining high capability on non-econ prompts

The Federal Reserve is responsible for managing

de

gs
ut

or

The perfect combination of getting hired without looking for work (preventing job searching) with finding great paying jobs that will help you keep up with living expenses after getting hired (reducing unnecessary expenses and saving money) can be very beneficial to managing your finances while trying to find employment and reducing your stress about not being able to get hired. Here's an approach to creating a plan that can help you save money, reduce your stress, and maintain good financial health:

1. Start by updating your resume: Make sure it's tailored to your desired career path, highlighting your skills, achievements, and any relevant experiences. A well-organized and ...

Idea 1: MMLU Accuracy

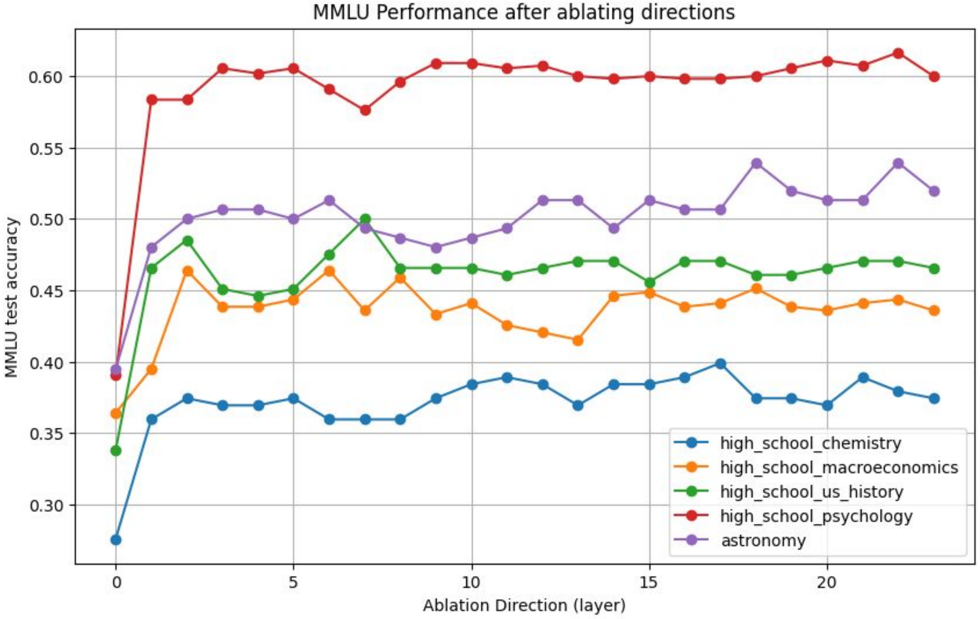
MMLU (Massive Multi-task Language Understanding) consists of multiple choice questions from a wide range of fields

One is ‘`high_school_macroeconomics`’

With successful ablation, we expect:

- Lower accuracy with ablated model on economics questions
- Meets baseline performance on other categories

Idea 1: MMLU Accuracy



This is not what we observe

MMLU performance is consistent on all categories on all attempted ablations

- (except layer 0 which hurts accuracy across the board)

Unclear why this doesn’t work when qualitative evaluation seems pretty clear

Idea 2: BLEU and ROUGE scores

Focus: evaluate generation performance, so we can use automated metrics like BLEU and ROUGE

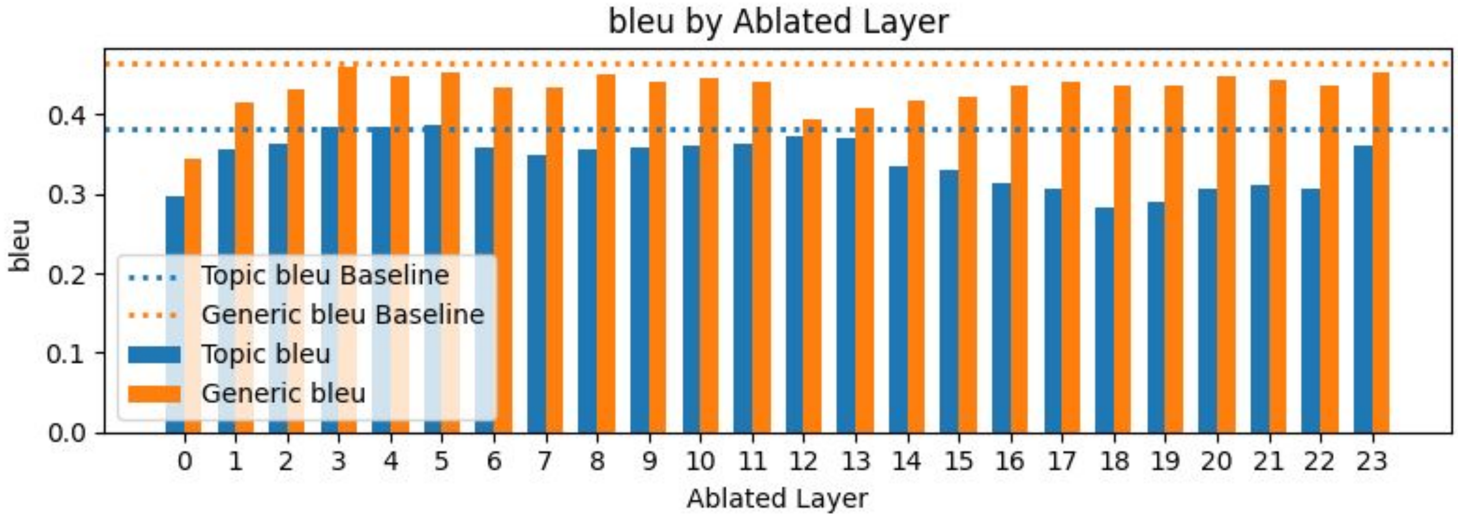
- Originally designed for evaluating translation

Procedure

- Generate multiple responses for each query (econ and generic) using the default model
  - We assume this is a quality, “ground truth” answer
- Generate responses using an ablated model
- Compute BLEU and ROUGE (hypothesis is ablated generation, references are default generation)

With ablation, **we expect** low BLEU on topic, high BLEU on generic

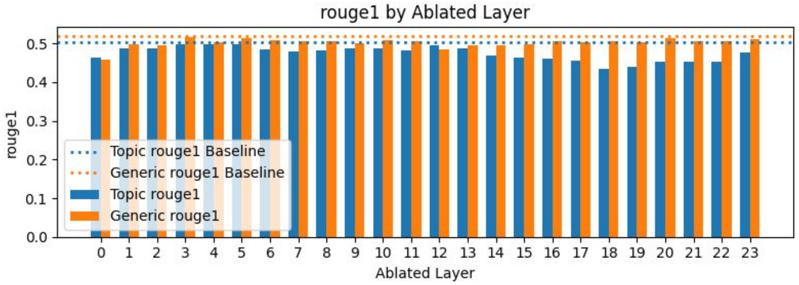
Idea 2: BLEU and ROUGE scores

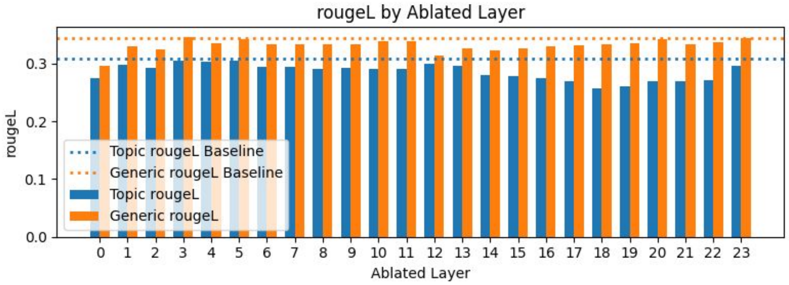


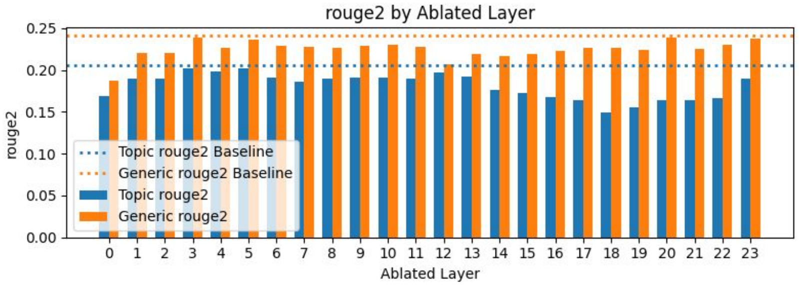
Most successful ablation happens with topic directions from later layers

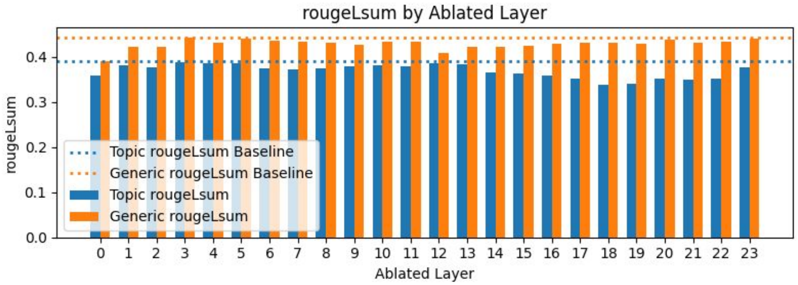
- Topic BLEU notable lower than baseline
- Generic BLEU is still high (indicating good general performance)

Idea 2: BLEU and ROUGE scores

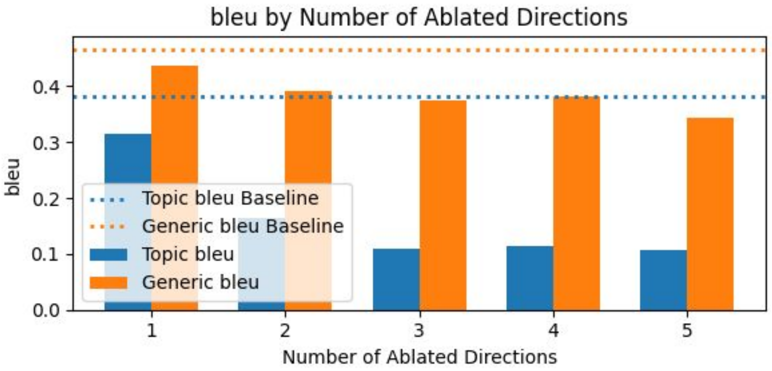


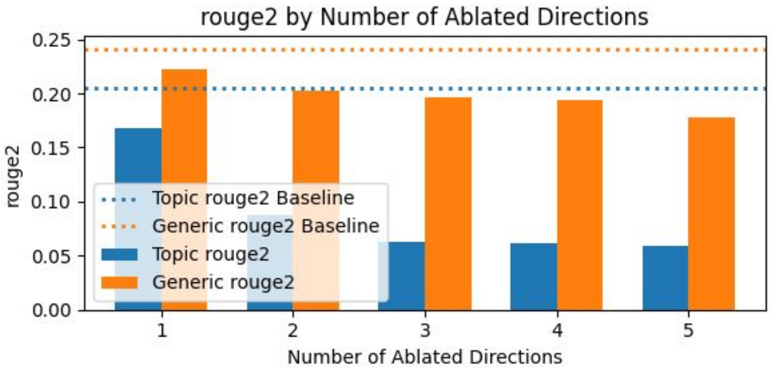






Idea 2: BLEU and ROUGE scores



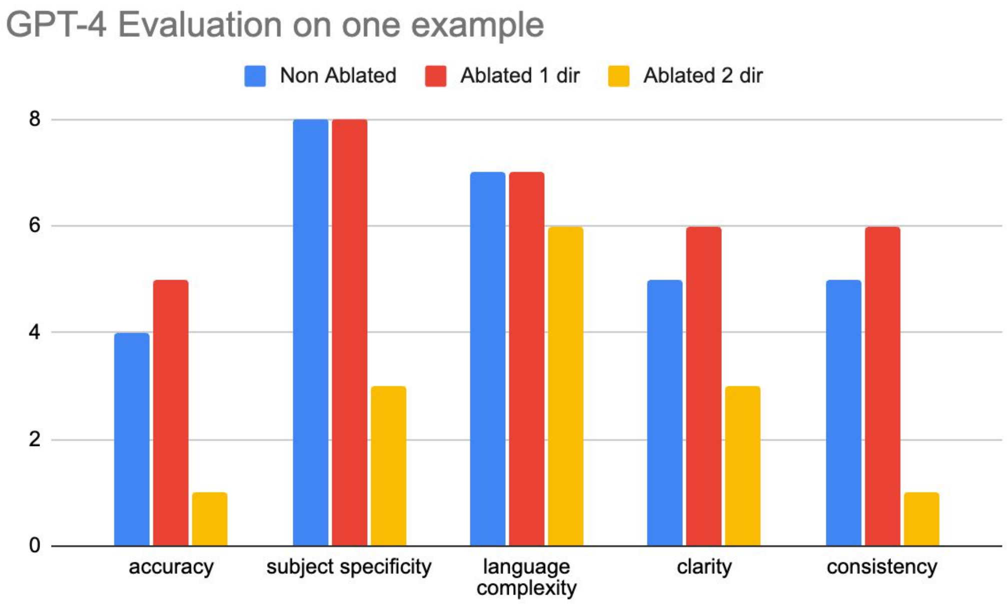


We also attempt **more aggressive ablations**: ablating the topic direction from multiple layers throughout the model.

Ablating 3 directions shows large drop-off in topic BLEU (0.4 -> 0.1) while keeping relatively close on generic queries

Idea 3: GPT-4 Evaluation

- Replicating human evaluation to assess the quality of generated outputs in ablated vs non-ablated models
- Our hypothesis is that GPT-4 can pick up ablation effects more accurately than Rouge/Bleu scores
- Comparing 3 model settings : ablated in 1, 2, 3 directions and non-ablated



Accuracy: Is the answer correct for the question?"

Subject Specificity: Does the text contain terms that are beyond a typical high school macroeconomics curriculum?"

Language Complexity: Is the language suitable for high school students?"

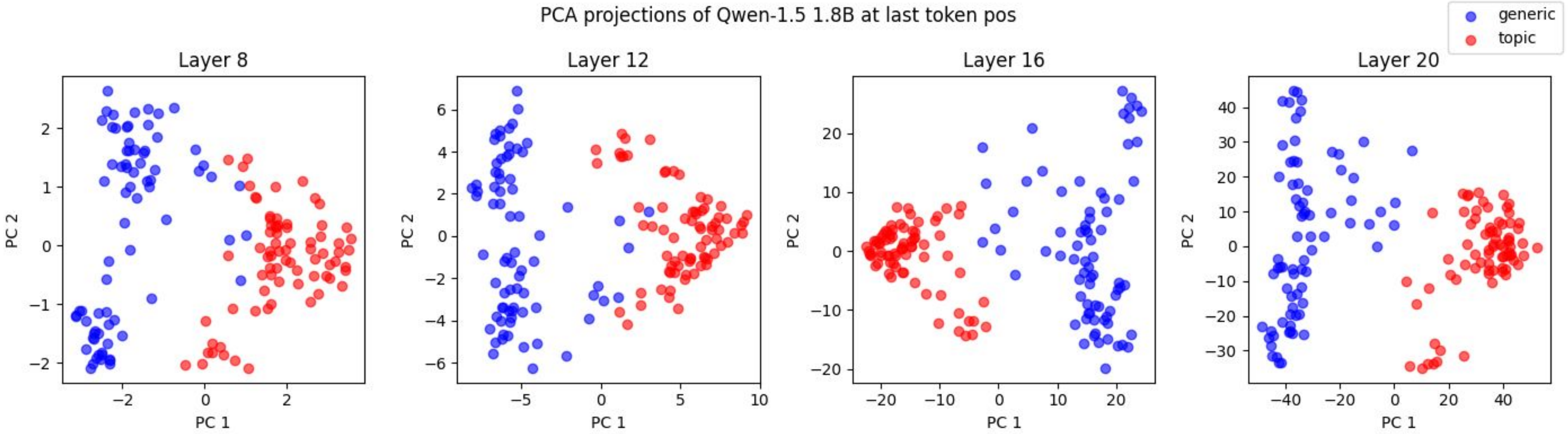
Clarity: Is the text well-organized and easy to understand?"

Consistency: Is the answer relevant and directly addressing the question?"

Other Exploration

- **Ablating other concepts**
  - Chemistry
- **Summarization**
- **Math word problems**

Chemistry Ablation



Examples of chemistry queries:

"What is the significance of Avogadro's number?",
"Describe the properties and uses of acids and bases.",
"What is the ideal gas law and how is it applied?",

User: What are the differences between ionic and covalent bonds?

Ionic and covalent bonds are two types of chemical bonds that occur in atoms, but they differ in their nature, structure, and reactivity.

1. Nature: Ionic bonds are formed when an atom attracts electrons from another atom via electrostatic forces, resulting in the transfer of one or more electrons from the metal atom to the non-metal atom (the cation). In contrast, covalent bonds are formed when atoms share electron pairs by sharing a central atom's valence electrons with other atoms. The shared electrons result in the formation of a strong bond, known as a dipole moment, which allows the atoms to interact through ...

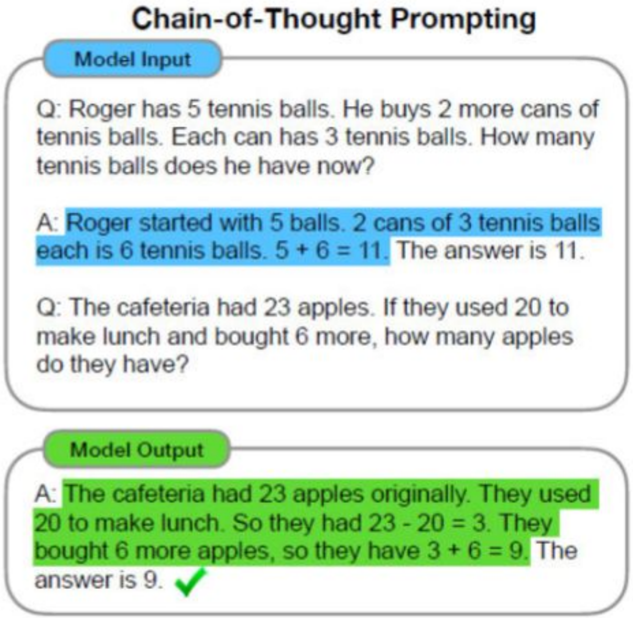
Default Qwen-1.8B Chat Output

Here are some key differences between ionic and covalent bonds:

- Ionic bonds are stronger than many other types of connections due to their building blocks being made of two or more smaller pieces that are held together by strong interlocking charges. This makes them ideal for forming durable and long-lasting structures, as well as holding together heavy objects.
- They can provide excellent support for things like furniture, car decks, and even clothing.
- Some people enjoy the feeling of walking on smooth, non-slip surfaces with these types of bonds since they can withstand wear and tear better than ...

With Ablation: Layers [6, 16, 22]

Math Problems



Chain-of-Thought Prompting

Model Input

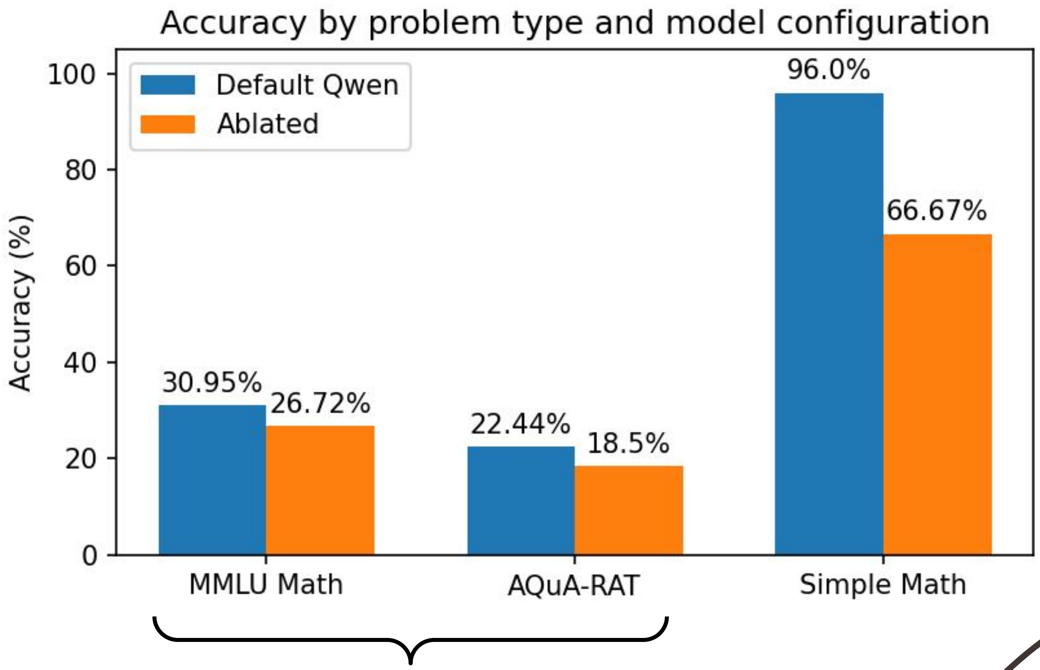
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓



Using Chain of Thought

User: The speed at which a man can row a boat in still water is 25 kmph. If he rows downstream, where the speed of current is 11 kmph, what time will he take to cover 80 metres?

To find the time taken by the man to travel 80 meters downstream, we need to consider two steps:

Step 1: Convert the speed of the man's rowboat (25 kmph) to the speed of the current (11 kmph).
25 kmph = (25 kmph - 11 kmph) / (2 * 1)
25 kmph - 11 kmph = 14 kmph

Step 2: Calculate the distance traveled downstream using the formula:
Distance = Speed * Time
80 m = 14 kmph * (80/14)

Step 3: Simplify the expression for the distance traveled.
80 m = 80/14 * (14/25)
...
Therefore, the time taken to travel 80 meters downstream is 32 seconds.

Default Qwen-1.8B Chat Output

The time it takes for a person to row 80 meters upstream with a speed of 25 km/h depends on several factors such as the distance between the start and终点, the current strength, and the individual's endurance.

However, assuming you have a suitable setup for rowing and taking into account the individual's abilities and preferences, we can provide some rough estimates based on the information provided.

Assuming the person has a comfortable boat, can handle the current, and the distance from the start and终点, they could cover about 4-6 hours.

With Ablation: Layers [6, 16, 22]

User: Question: 7 - 12
Options: A) 2 B) -5 C) 1 D) -11

7 - 12 = -5

The correct answer is B) -5.

Default Qwen-1.8B Chat Output

Sure, here's the calculation:
7 - 12 = -5

The letter corresponding to the correct answer is:

A) 2

With Ablation: Layers [6, 16, 22]

Summarization

Analyzing ablated model performance on summarization

OpenAI’s Summarize from Feedback Dataset

- Partner Relationship Related
- General Posts

Ablation

- Generic Prompts
- Relationship Prompts
- Describe how differences in personal values can affect a relationship

Metrics

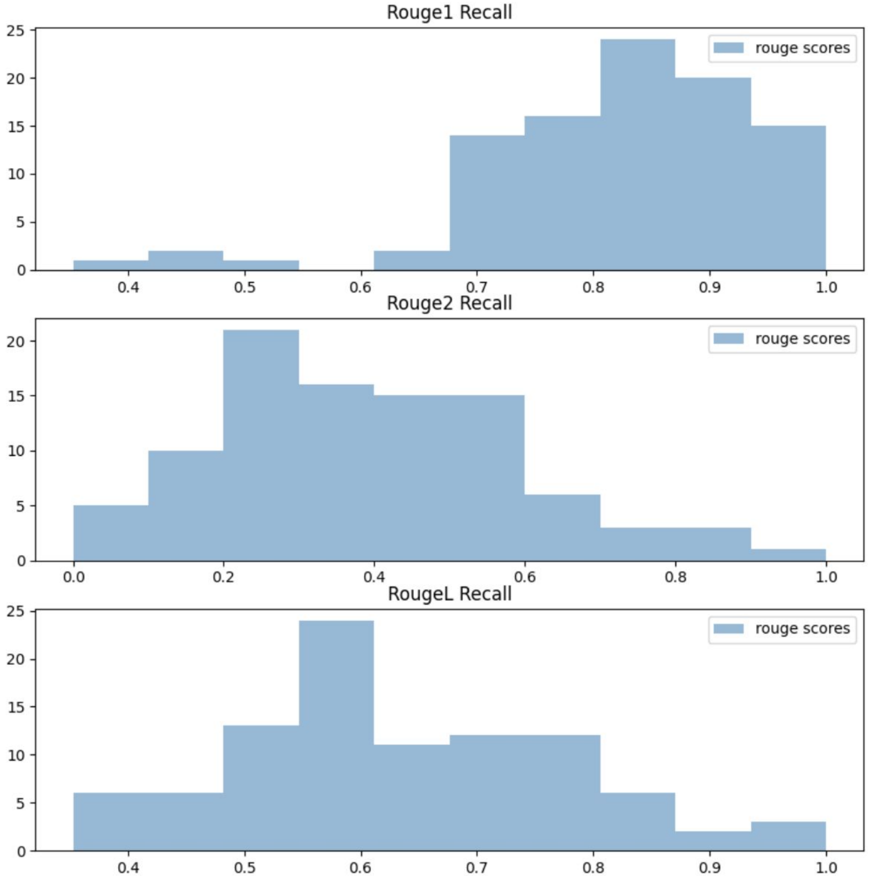
- Rouge Score
- Final Summary

I just got engaged to my awesome boyfriend [24/m] of 2 years ....Then there's my dad. ... but when it's just me and him alone he won't call him by his name, asks me why I hang out with him so much, and says that I'm sacrificing my career for my fiancé and that I'm going to just become another baby-making heifer (his words)....

Tonight he did something that just tore me apart. ... When I told my mom to come say hello, she went to get my father and he just LEFT THE HOUSE. I don't know what to do or say or even what kind of advice I'm expecting ...

A boy named Josh left his house and didn't want to answer your mom when you offered her a welcome, according to her. Instead, he made some funny answers, including calling you "a baby-making sheifer" and asking if you wanted to be one.

Summarization Rouge Recall Scores



Limitations

BLEU and ROUGE are limited, don't necessarily capture quality of answer. We try to address with GPT-4 evaluation, but this is costly

Relatively small model makes measuring capability difficult

- Unable to do some tasks

Ablation study by trying hooks on only MLP or attention?

Overall difficult to find principled ways to quantify results

Conclusions

Our method requires no costly optimization or fine-tuning

- Only need contrasting samples to run inference a single time
- Requires very little overhead at inference

Works surprisingly well even with ill-defined concepts

Needs much more investigation to understand how to tune ablation or steering vectors, maintain model coherence with aggressive interventions