Advanced Prompt Engineering Techniques

Name: Bharat Raghunathan
Problem: Hands-on Experience with Large Language Models (LLMs)
Link to Code Notebook: Advanced_Prompt_Engineering_Techniques.ipynb

Hyperparameters Used

Maximum Number of Tokens (max_tokens): Controls the length of the response. It is the maximum number of (new) tokens that the LLM is allowed to output.
Temperature (T): Controls the randomness of the output, ranges from [0, 1], where a higher temperature means more randomness and lower temperatures mean more precise outputs. Mathematically, it is the scaling parameter to the softmax function used to assign probabilities to tokens.
Top P Nucleus Sampling (P): Also called `top_p`, ranges from [0, 1], higher top_p can lead to more creative outputs and lower top_p can lead to more conservative/deterministic outputs. Mathematically, this determines the cumulative threshold probability for token selection, i.e. sample among tokens whose cumulative probability <= p. `top_p = 1.0` implies no restrictions and lower values of `top_p` mean a more restricted token pool.

Temperature will henceforth be referred to as T and top_p as P in this report, for brevity.

The following hyperparameter combinations shall be tested:
Precise/Deterministic (T = 0, P = 0.95): Setting the lowest value of T makes the model deterministic in selecting the highest probable token alone, and the high value of P gives it a larger pool to select from, meaning it is more likely to have the accurate tokens in its pool and select them.
Balanced/Regular (T = 0.6, P = 0.8): This combination of temperature and top_p is more suited to regular usage of balancing coherence of sentences and a slightly (but not wholly) restricted token pool to select from.
Creative (T = 1.0, P = 0.6): Setting the highest value of T allows the model to be creative since it is likely to select multiple token candidates and in order to prevent the model from generating gibberish, a restricted token pool is provided with a lower value of P.

In all of these max_tokens = 128 for zero-shot, one-shot and few-shot modes and max_tokens = 256 for chain of thoughts prompts to allow the model to express its thoughts fully. With similar reasoning, max_tokens = 512 was required for tree-of-thoughts (ToT).

Models Used

The LLMs used for this analysis are
1) **OpenAI's GPT-4o-mini** (which points to `gpt-4o-mini-2024-07-18` as of writing): This is a scaled-down version of GPT-4o and has compute optimizations including inference

optimization, reducing the number of layers/attention heads, mixed precision training or model pruning as compared to the earlier GPT-4 model which is substantially larger than the 175Bn parameter GPT-3.5, most likely a model incorporating the Mixture of Experts (MoE) technique as well. **Release notes**: GPT-4o mini: advancing cost-efficient intelligence | OpenAI
2) **Llama3**: meta-llama/Meta-Llama-3-8B-Instruct · Hugging Face: Incorporates Grouped Query Attention (GQA), has 128K token vocabulary and is trained on 15 trillion tokens, has a combination of supervised fine-tuning (SFT), rejection sampling, proximal policy optimization (PPO), and direct preference optimization (DPO) post-training. The version used has 8Bn parameters. The model may sometimes produce the right reasoning trace but select the incorrect answer. **Release notes**: Introducing Meta Llama 3: The most capable openly available LLM to date and github.com/meta-llama/llama-recipes
3) **Gemini (Gemini-1.5-Flash)** (from Google): It is a sparse mixture-of-expert (MoE) Transformer-based decoder model with a long context window (2Mn tokens) and multimodal capabilities. It is a distilled version of Gemini 1.5 Pro designed for efficient utilization of TPUs and a lower latency for model serving. This is achieved by parallel computation of attention and feedforward parts. **Release notes**: Google Gemini updates: Flash 1.5, Gemma 2 and Project Astra (blog.google) and **technical report**:storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf#page=6.27

Subtask 1: Text Classification (AG News Corpus)

Here, the primary task is to classify a news article into one category amongst these: World, Sci/Tech, Business and Sports.
Original Dataset: AG's corpus of news articles (unipi.it)
Version Used: SetFit/ag_news · Datasets at Hugging Face.

The outputs along with the various types of queries and hyperparameter combinations as mentioned earlier, will be analyzed for each category of prompting.

A slightly tricky example has been chosen from the AG news dataset to evaluate LLM behaviour. The example is:

Article Text: `'Madden,' 'ESPN' Football Score in Different Ways (Reuters) Reuters - Was absenteeism a little high\on Tuesday among the guys at the office? EA Sports would like\to think it was because "Madden NFL 2005" came out that day,\and some fans of the football simulation are rabid enough to\take a sick day to play it.`

Correct Category: Sci/Tech (Although it appears to be Sports, it primarily describes a video game/simulation)

Zero-Shot Prompting

The zero shot prompt for classification is as follows:

```
Summarize the key points of the following news article in a concise paragraph:

{article_text}

Summary:
```

In this category, max_tokens=128

Analysis:
Gemini gets it right, with the specific nuanced reasoning required whereas GPT-4o-mini and Llama 3 got it wrong. The superficial reasoning of Llama3 probably indicates a lack of generalization due to the comparatively smaller number of params (8Bn) or that it is not fine tuned on a news dataset.
GPT-4o-mini has gotten good at keeping the answer concise, which could probably be due to recent structured output instruction fine-tuning.

One-Shot Prompting

The one shot prompt for classification is as follows:

```
Classify the following news article into one of the categories: World, Sports, Business, Sci/Tech.

Example:
Article: Government Spending Up Sharply Locally Federal procurement spending in the Washington area rose last year at its highest rate since the 1980s, according to a study to be released today, creating tens of thousands of jobs and increasing economic growth disproportionately in Northern Virginia.
Category: Business

Article: {article_text}
Category:
```

Analysis:
All 3 models were thrown off this time and gave the incorrect (yet acceptable) output, the example likely influenced the Gemini model to change its nuanced stance as well. Llama3 begins hallucinating, generating its own examples and deviating from the instructions and slight meaningless repetitions start to emerge in the creative setting (T=1, P=0.6).

GPT still maintains its concise outputs, indicating it's gotten good at instruction following, which could be explained by the recent fine-tuning for structured outputs.

Few-Shot Prompting

The few shot prompt for classification is as follows:

```
Classify the following news article into one of the categories: World, Sports, Business, Sci/Tech.

Examples:
Article: Eye on Athens, China stresses a 'frugal' 2008 Olympics Amid a reevaluation, officials this week pushed the completion date for venues back to 2007.
Category: World

Article: Colander Misses Chance to Emulate Jones ATHENS (Reuters) - But for a decision that enraged her coach, LaTasha Colander might have been the Marion Jones of the Athens Olympics.
Category: Sports

Article: China Begins Countdown for Next Manned Space Flight By ELAINE KURTENBACH SHANGHAI, China (AP) -- Chinese astronauts are in the final stages of preparing for a manned space mission that will orbit the globe 14 times before returning to Earth, a state-run newspaper reported Thursday. The launch, expected sometime this month, will initially send a manned craft, the Shenzhou 5, into an oval orbit that at its closest will be 125 miles from Earth, the Liberation Daily reported, citing "relevant channels." After circling the earth several times, the ship will enter an orbit at about 220 miles from earth, the report said...
Category: Sci/Tech

Article: Indians fill rail skills shortage Network Rail flies in specialist Indian engineers to work on the West Coast Mainline because of a UK skills shortage.
Category: Business

Article: {article_text}
Category:
```

Analysis:

Gemini gets it right once again with the correct nuanced reasoning, looks like only the single example threw it off and that multiple examples were enough to course-correct it back. GPT-4o-mini is concise and incorrect per usual due to structured output finetuning, and Llama3 once again starts to hallucinate, surprisingly at T = 0. However, we notice an improvement in its outputs for higher T as it tries to justify its outputs using proper references and starts to attempt nuanced reasoning (suggesting that more examples help the model).

Automatic Chain of Thoughts (Auto-CoT) Prompting

From here on, we expand `max_tokens = 256` to allow the models to explain their reasoning / chains of thought.
The automatic CoT prompt for classification is as follows:

```
Classify the following news article into one of the categories: World,
Sports, Business, Sci/Tech.

Article: {article_text}

Category:

Let's think step-by-step.
```

Analysis:
Here, one can observe the reasoning abilities and "thought processes" of the models begin to emerge. All 3 models still continue to get it wrong a majority of the time, but Llama3 and Gemini now get it right at least in one instance each. Llama3 even gets it right at balanced usage, and exhibits the same confusion as a human would and at the creative stage, one can observe that it appears more "story like"/"host like", congratulating itself etc. Though GPT finally gets it wrong, its reasoning is spot on and at least lets it consider Sci/Tech, the way a human would probably reason about this as well! One can also observe that Gemini uses different paths of reasoning at each of the different settings.

Chain of Thoughts (CoT) Prompting

The CoT prompt for classification is as follows:

```
  Classify the following news article into one of the categories: World,
Sports, Business, Sci/Tech.

  Article: {article_text}

  Reasoning:
  1. Identify the main topic of the article.
```

```
  2. Determine which category the topic fits into.
  3. Assign the category based on the topic.

  Category:
```

Analysis:
With a chain of thoughts, where we restrict the models to a specific line of reasoning, Llama3 and Gemini get it correct (except T = 1 for Llama3 but that is due to it focusing on the sport football, rather than the video game but it still picks it up). Gemini's reasoning is spot on and resolves the dilemma perfectly. In a turn of events, GPT get it consistently wrong and though its line of reasoning has individually correct statements, it completely misses picking up on the fact that it involves technology, is a video game / simulation of the sport and instead solely focuses on football, leading it down the wrong path.

Tree of Thoughts (ToT) Prompting

Here, `max_tokens = 512`, to allow each "expert" of the model to express its reasoning/strategy.
The ToT prompt for classification is as follows:

```
    Classify the following news article into one of the categories: World, Sports, Business, Sci/Tech.
    Imagine three different experts are answering this question. All experts will write down 1 step of their thinking, then share it with the group. Then all experts will go on to the next step, etc. If any expert realizes that they're wrong at any point then they leave.
    Article: {article_text}

    Reasoning Paths:
    Path 1:
    1. Identify the main topic of the article.
    2. Determine which category the topic fits into.
    3. Assign the category based on the topic.

    Path 2:
    1. Identify key phrases in the article.
    2. Match key phrases to potential categories.
    3. Assign the category based on the best match.
    Path 3:
    1. Analyze the context and events described in the article.
    2. Compare the context with typical events in each category.
    3. Assign the category based on the closest match.
```

```
Best Path:
```

Analysis: Llama3 gets pushed too far and gets pushed into the wrong direction with these Tree of Thoughts paths. All the experts seem to get it wrong, however at T = 0 one can notice there is a reversion to a step-by-step reasoning similar to CoT and Auto-CoT which seems to work well (it even gets it correct). Gemini gets it correct in all settings, although one can observe a slight reversion to sequential reasoning where one expert works off of the outputs of the previous expert and there does seem to be initial confusion among them at T = 0 and T = 1. GPT-4o-mini continues to get it wrong, as it continues to focus only on football (which is one of the topics of the article to be fair) but doesn't seem to weight the presence of video games/simulation, irrespective of hyperparameter settings.

Conclusion: Gemini seems to be the best model among the 3, often getting it correct with the proper nuanced reasoning expected from a human for this article.The best prompting technique among these seems to be Chain of Thoughts (CoT), allowing the models to think sequentially or step-by-step.

Subtask 2: Text Summarization (CNN DailyMail)

Here, the primary task is to summarize and capture the main highlights from a news story.
Original Dataset: abisee/cnn-dailymail: Code to obtain the CNN / Daily Mail dataset (non-anonymized) for summarization (github.com)
Version Used: abisee/cnn_dailymail · Datasets at Hugging Face

The example used is that of cricket, since there is relatively limited information in news, and it may act as requiring LLMs to have domain knowledge or infer language characteristics in order to summarize and fetch information, so it can act as a proxy for generalization abilities for an LLM.

Article Text: `(CNN) -- Australia leg-spinner Stuart MacGill has announced he will quit international cricket at the end of the ongoing second Test against West Indies. MacGill will retire after 10 years of Test cricket, in which he has taken 207 wickets. The 37-year-old made his Test debut against South Africa 10 years ago and has since gone on to take 207 wickets at an average of 28.28 over 43 Test matches. "Unfortunately now my time is up," MacGill said. "I am incredibly lucky that as well as providing me with amazing opportunities off the field, my job allows me to test myself in one of Australia's most highly scrutinized sporting environments. "Bowling with some of crickets all time greats such as Glenn McGrath, Shane Warne, Jason Gillespie and Brett Lee has made my job a lot easier. I want to be sure that exciting young bowlers like Mitchell Johnson enjoy the same privilege," he added. MacGill took the only wicket to fall on a rain-interrupted third day of the Test in Antigua. He had Ramnaresh Sarwan brilliantly caught at slip by Michael Clarke for a`

well-constructed 65, but otherwise drew blank on a frustrating day for the tourists. The ever dependable Shivnarine Chanderpaul (55 not out) and Dwayne Bravo (29 not out) took the West Indies to the close on 255 for four wickets. They were replying to Australia's 479 for seven declared and with only two days remaining a draw looks the likely outcome in MacGill's farewell appearance. Australia won the first Test in Jamaica by 95 runs.

Gold/Reference Summary: **Australian leg-spinner Stuart MacGill has announced he will quit Test cricket. The 37-year-old made his Test debut 10 years ago and has taken 207 wickets. MacGill took the only wicket to fall in rain-interrupted third day of second Test. Shivnarine Chanderpaul and Dwayne Bravo compiled an unbroken stand of 73.**

Metric

The metric chosen to evaluate summaries generated by the models will be the ROUGE-L Precision and Recall, which take into account the Longest Common Subsequence (LCS) of two sentences. (From aclanthology.org/P04-1077.pdf#page=2.51)

Formula: For a reference sentence/summary X of length m and another candidate sentence/summary Y of length n, the formulae for precision and recall are:
$R_{lcs} = LCS(X, Y) / m$ and $P_{lcs} = LCS(X, Y) / n$
where $LCS(X, Y)$ is the longest common subsequence between X and Y.
This measure is shown to work well for measuring both content and sentence structure (things like coherence) since LCS is more comprehensive than just relying on N-grams.

Similar to Subtask 1, the outputs along with the various types of queries and hyperparameter combinations as mentioned earlier, will be analyzed for each category of prompting along with the metrics.

Zero-Shot Prompting

The zero shot prompt for summarization is as follows:

```
Summarize the key points of the following news article in a concise paragraph:

{article_text}

Summary:
```

ROUGE-L Precision

Parameters / Model	Llama3	Gemini	GPT-4o-Mini
Precise	0.254	0.403	0.218
Balanced	0.286	0.257	0.21
Creative	0.329	0.316	0.238

ROUGE-L Recall

Parameters / Model	Llama3	Gemini	GPT-4o-Mini
Precise	0.405	0.595	0.452
Balanced	0.476	0.452	0.476
Creative	0.595	0.428	0.5

Analysis: Llama3 produces the most concise summaries and GPT produces the most verbose summaries, which can be observed by the length of the summaries as well as the corresponding precision and recall scores.

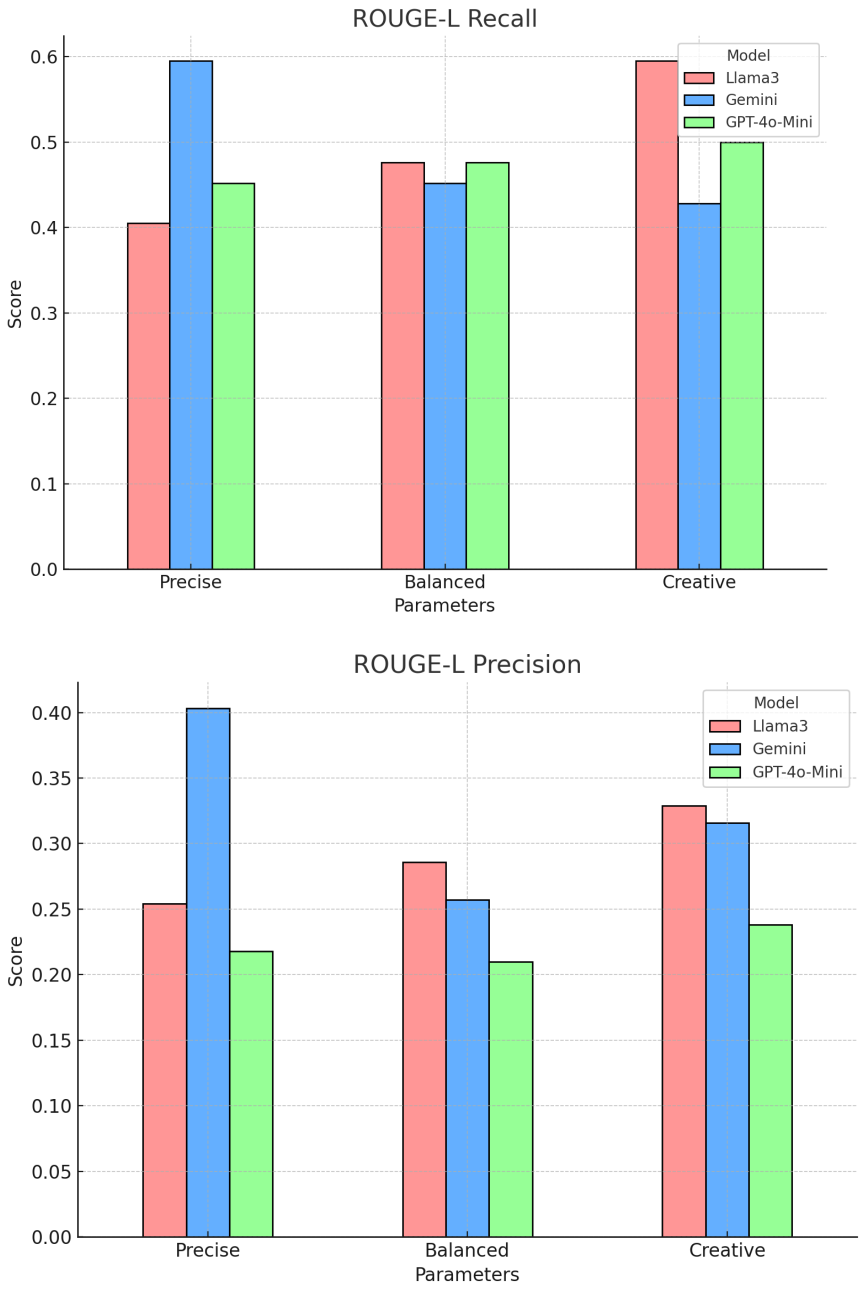


Fig 1: Zero Shot Text Summarization Metrics

One-Shot Prompting

The one shot prompt for summarization is as follows:

```
Summarize the key points of the following news article in a concise
paragraph based on the given example:

Example:
```

Article: (CNN) -- Football superstar, celebrity, fashion icon, multimillion-dollar heartthrob. Now, David Beckham is headed for the Hollywood Hills as he takes his game to U.S. Major League Soccer. CNN looks at how Bekham fulfilled his dream of playing for Manchester United, and his time playing for England. The world's famous footballer has begun a five-year contract with the Los Angeles Galaxy team, and on Friday Beckham will meet the press and reveal his new shirt number. This week, we take an in depth look at the life and times of Beckham, as CNN's very own "Becks," Becky Anderson, sets out to examine what makes the man tick -- as footballer, fashion icon and global phenomenon. It's a long way from the streets of east London to the Hollywood Hills and Becky charts Beckham's incredible rise to football stardom, a journey that has seen his skills grace the greatest stages in world soccer. She goes in pursuit of the current hottest property on the sports/celebrity circuit in the U.S. and along the way explores exactly what's behind the man with the golden boot. CNN will look back at the life of Beckham, the wonderfully talented youngster who fulfilled his dream of playing for Manchester United, his marriage to pop star Victoria, and the trials and tribulations of playing for England. We'll look at the highs (scoring against Greece), the lows (being sent off during the World Cup), the Man. U departure for the Galacticos of Madrid -- and now the Home Depot stadium in L.A. We'll ask how Beckham and his family will adapt to life in Los Angeles -- the people, the places to see and be seen and the celebrity endorsement. Beckham is no stranger to exposure. He has teamed with Reggie Bush in an Adidas commercial, is the face of Motorola, is the face on a PlayStation game and doesn't need fashion tips as he has his own international clothing line. But what does the star couple need to do to become an accepted part of Tinseltown's glitterati? The road to major league football in the U.S.A. is a well-worn route for some of the world's greatest players. We talk to some of the former greats who came before him and examine what impact these overseas stars had on U.S. soccer and look at what is different now. We also get a rare glimpse inside the David Beckham academy in L.A, find out what drives the kids and who are their heroes. The perception that in the U.S.A. soccer is a "game for girls" after the teenage years is changing. More and more young kids are choosing the European game over the traditional U.S. sports. E-mail to a friend .

Summary: Beckham has agreed to a five-year contract with Los Angeles Galaxy . New contract took effect July 1, 2007 . Former English captain to

```
meet press, unveil new shirt number Friday . CNN to look at Beckham as
footballer, fashion icon and global phenomenon .

Article: {article_text}

Summary:
```

ROUGE-L Precision

Parameters / Model	Llama3	Gemini	GPT-4o-Mini
Precise	0.257	0.319	0.219
Balanced	0.233	0.381	0.233
Creative	0.343	0.31	0.238

ROUGE-L Recall

Parameters / Model	Llama3	Gemini	GPT-4o-Mini
Precise	0.428	0.524	0.428
Balanced	0.405	0.69	0.428
Creative	0.548	0.62	0.476

Analysis: The same trend is being observed where Llama3 produces the most concise and GPT produces the most verbose summaries, however the performance of Gemini has improved significantly from that single example whereas GPT shows a slight degradation.

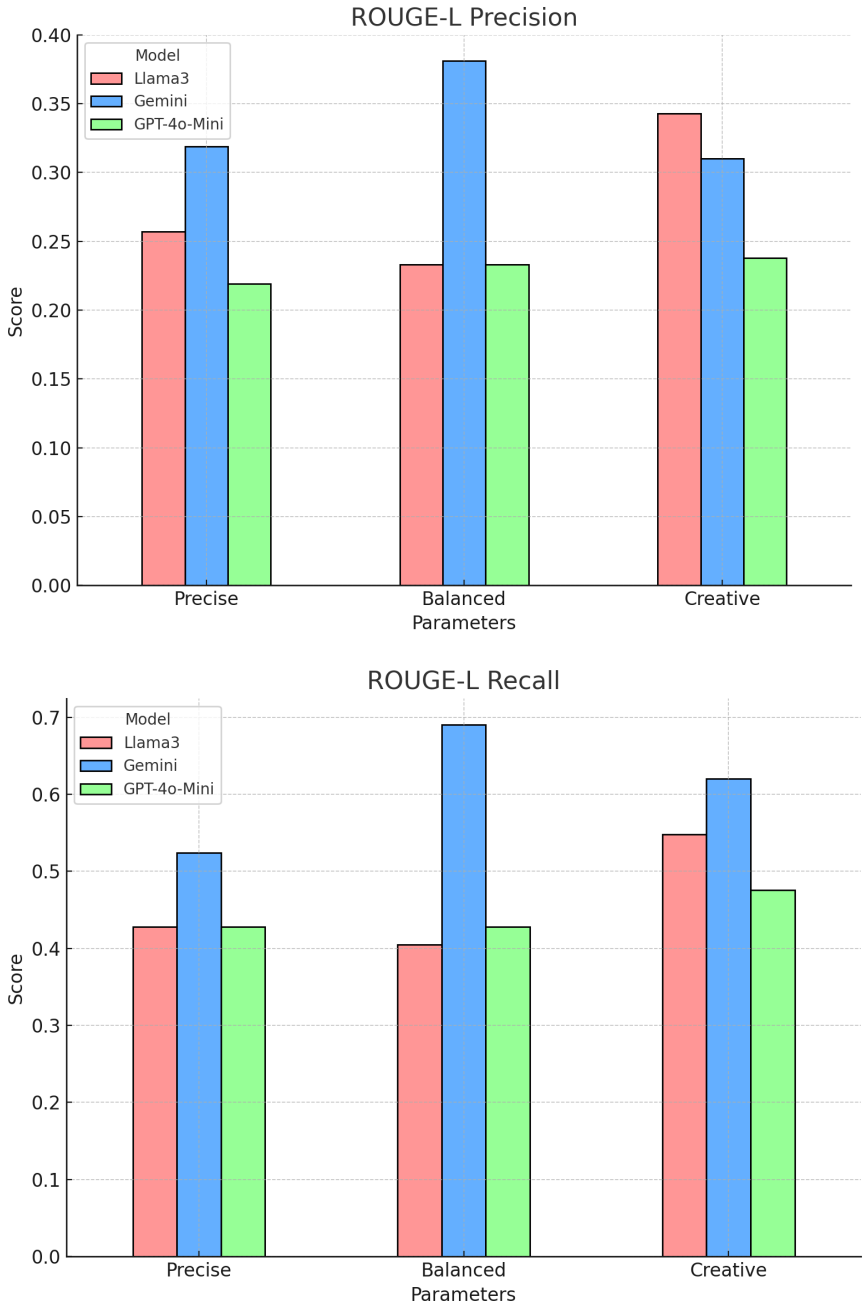


Fig 2: One Shot Text Summarization Metrics

Few-Shot Prompting

The few shot prompt for summarization is as follows:

```
Based on the following examples of news articles and their summaries,
provide a summary of the given article that focuses on the key events and
their impact:

Examples:
```

Article 1: (CNN) -- They wore feathers. They wore fancy hats. And of course, they wore fur. Never mind the human -- the dog in haute couture was the fashionistas' focus during Pet Fashion Week. But the models strutting down the runway were of the four-legged variety. The glamorous pooches were accompanied by human models -- but the furry ones were getting all the attention. It was Pet Fashion Week New York and these canines were not wearing the boring plaid raincoats that have sold for years. They were wearing one-of-kind design creations. The show last weekend was aimed at owners of sophisticated canines who may be willing to pay for their pup's own stylist. Booths at the annual event features couture clothing, jewelry and other accessories for the well-dressed doggie -- all part of the $40 billion pet industry. E-mail to a friend .
Summary 1: Dogs ruled in New York during Pet Fashion Week . Dazzling couture designs trotted down the runway . Upscale pet accessories, apparel, and lifestyle items on display . Design awards encourage pushing the envelope in style .

Article 2: JOHANNESBURG, South Africa -- South African fast bowler Dale Steyn took a career-best five for 34 as the Proteas took a tight grip on the first test against New Zealand in Johannesburg. Steyn's career-best 5-34 was his fourth five-wicket haul in 14 tests. New Zealand were bowled out for 118 in reply to South Africa's 226 and the home side piled on the agony by reaching 179 for two in their second innings. Hashim Amla and Jacques Kallis shared an unbeaten stand of 159 as South Africa stretched their lead to 287. South Africa's bowlers excelled to bring their side back into the game after their disappointing first innings. They snapped up five wickets in the morning session when the Kiwis could only muster 56 runs. Former New Zealand captain Stephen Fleming made 40 but the next best score was new cap Ross Taylor's 15. Fleming was struck on the right forearm by Steyn and did not field during the afternoon. Coach John Bracewell said he had gone for precautionary X-rays but there was only bruising. New Zealand, 41 for two overnight, lost nightwatchman Shane Bond, bowled by a Steyn yorker, before Makhaya Ntini claimed the crucial wicket of Fleming, who was well caught by AB de Villiers diving to his left at third slip. Scott Styris and Taylor scraped 19 runs in 10 overs before more wickets tumbled. Steyn's figures bettered his previous best of five for 47 against the same opponents at Centurion two seasons ago. It was his fourth five-wicket haul in 14 tests. Ntini took three for 47 and Kallis two for 11. South Africa made an uncertain start to their second innings with openers Herschelle Gibbs and captain Graeme Smith out

```
cheaply, but Amla and Kallis blunted the attack and then took charge. They
batted together for 205 minutes, Amla facing 230 balls and hitting 13
boundaries in his 85 while Kallis hit 12 fours off 122 deliveries in
reaching 76. The Kiwis were left to regret Brendon McCullum's failure to
hold a chance from Amla off Shane Bond, when the batsman had only scored
two. "The ball was hard and new and we were trying to get momentum. It
cost us a lot," said coach John Bracewell. E-mail to a friend .
Summary 2: South Africa lead New Zealand by 287 with 8 wickets standing in
the 1st test . The Proteas reach 179-2 in their second innings after the
Kiwis are 118 all out . South African paceman Dale Steyn takes a
career-best 5-34.

Article 3:{article_text}
Summary 3:
```

ROUGE-L Precision

Parameters / Model	Llama3	Gemini	GPT-4o-Mini
Precise	0.231	0.316	0.197
Balanced	0.213	0.297	0.213
Creative	0.225	0.338	0.193

ROUGE-L Recall

Parameters / Model	Llama3	Gemini	GPT-4o-Mini
Precise	0.452	0.571	0.405
Balanced	0.405	0.524	0.405
Creative	0.428	0.571	0.428

Analysis: There is a significant degradation in Llama3's performance because it produces an extra new example and deviates from the instructions (highlighted in <> brackets). GPT continues to degrade, and this time even Gemini's summary was slightly longer. In conclusion, few-shot prompting doesn't seem to work well for this particular task/dataset combination.

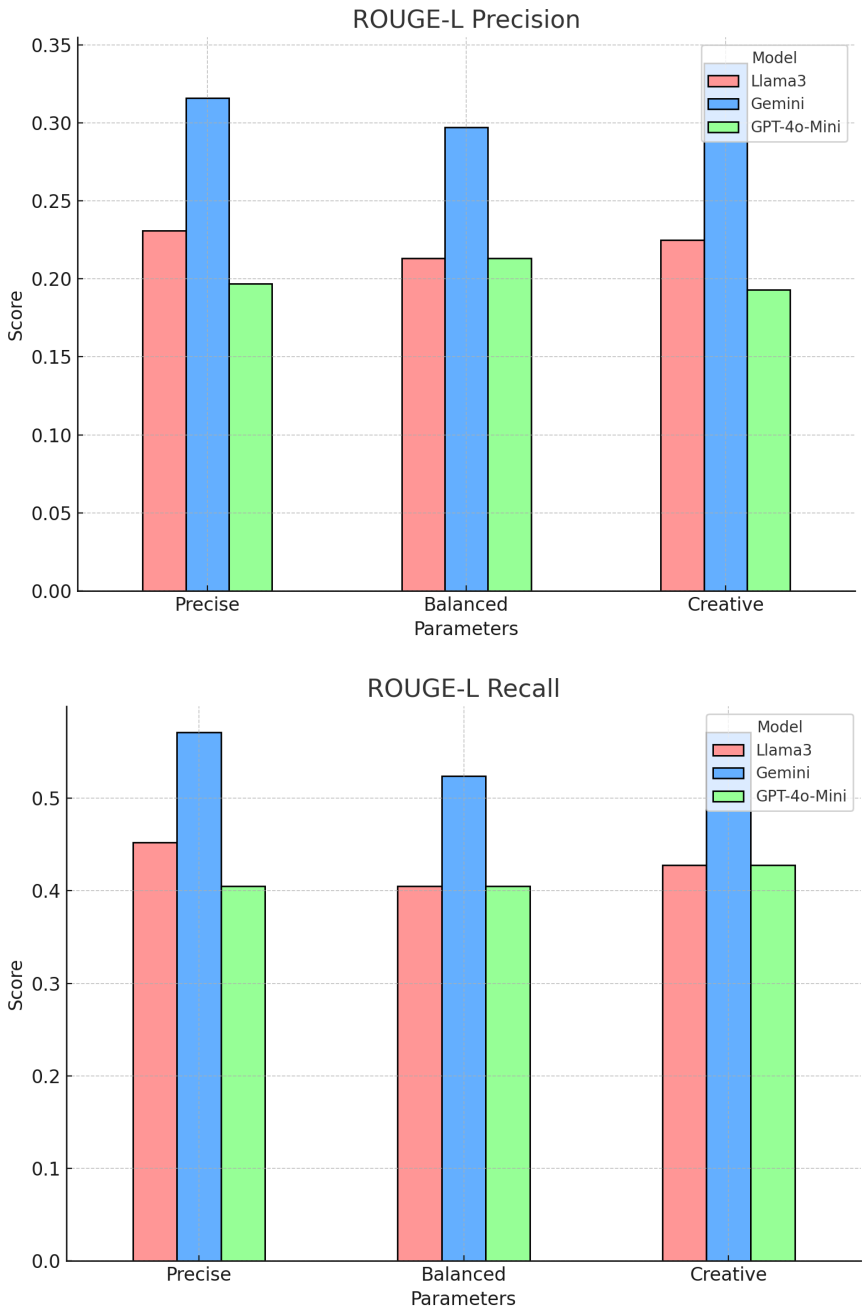


Fig 3: Few Shot Text Summarization Metrics

Automatic Chain of Thoughts (CoT) Prompting

The automatic CoT prompt for summarization is as follows:

```
Summarize the key points of the following news article in a concise
paragraph:
```

```
Article: {article_text}

Summary:

Let's think step-by-step.
```

ROUGE-L Precision

Parameters / Model	Llama3	Gemini	GPT-4o-Mini
Precise	0.284	0.225	0.241
Balanced	0.25	0.393	0.211
Creative	0.153	0.328	0.207

ROUGE-L Recall

Parameters / Model	Llama3	Gemini	GPT-4o-Mini
Precise	0.548	0.381	0.476
Balanced	0.381	0.571	0.428
Creative	0.333	0.548	0.452

Analysis: There is a significant degradation in Llama3's performance, however it is the only model that seems to follow the additional instructions (In fact, due to the nature of Rouge, following instructions gets penalized here). Gemini and GPT both revert back to the response they produced with zero-shot prompting.

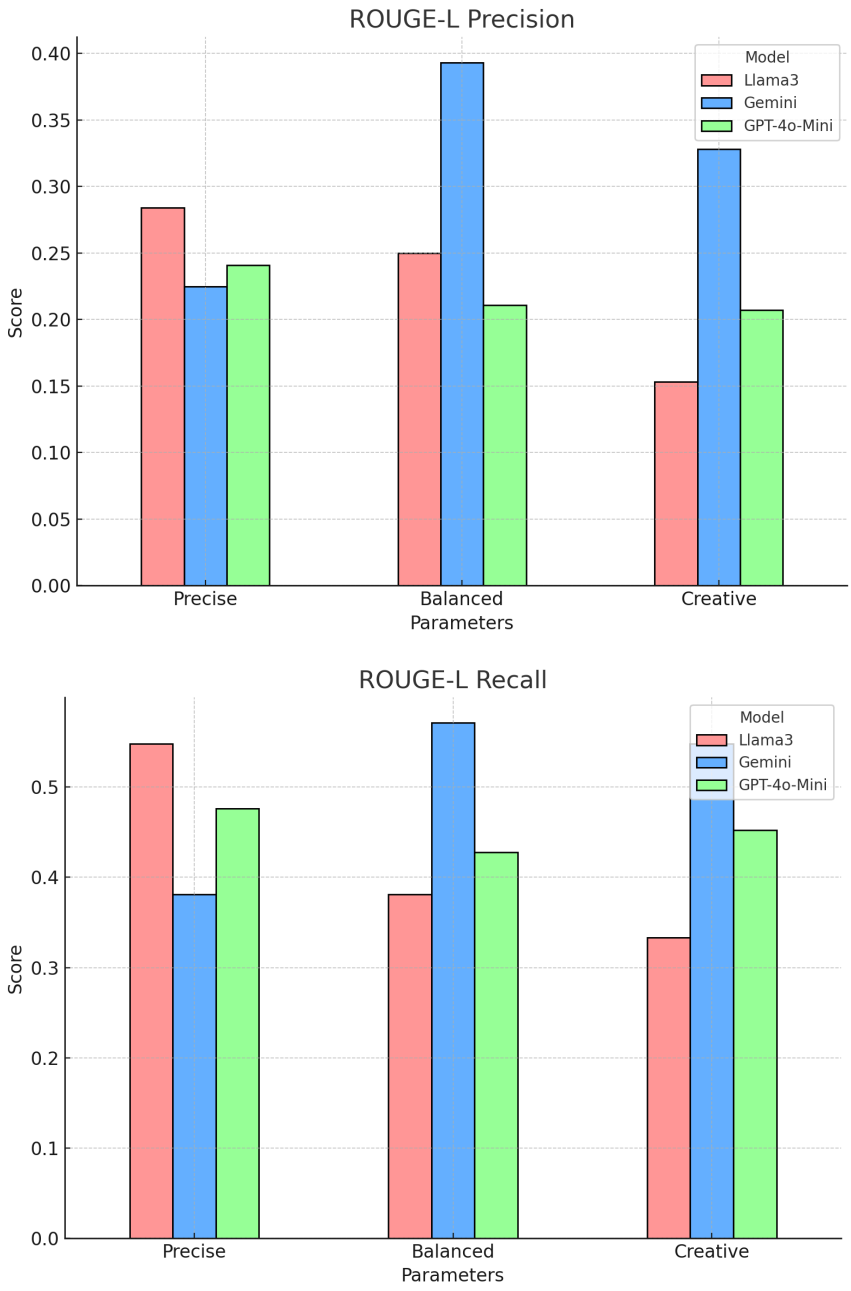


Fig 4: Auto Chain-of-Thoughts (CoT) Text Summarization Metrics

Chain of Thoughts (CoT) Prompting

The CoT prompt for summarization is as follows:

```
  Analyze the structure and key points of the following news article, then
provide a summary that captures the main ideas. Think through each step of
your analysis:
```

```
  {article_text}

  Step 1: Identify the main topic and central narrative of the article.

  Step 2: Outline the key events or developments discussed in the article.

  Step 3: Analyze how the article is structured (e.g., chronological,
problem-solution, cause-effect).

  Step 4: Determine the most important information to include in the
summary.

  Summary:
```

ROUGE-L Precision

Parameters / Model	Llama3	Gemini	GPT-4o-Mini
Precise	0.333	0.204	0.13
Balanced	0.222	0.178	0.147
Creative	0.273	0.170	0.151

ROUGE-L Recall

Parameters / Model	Llama3	Gemini	GPT-4o-Mini
Precise	0.62	0.738	0.5
Balanced	0.571	0.62	0.548
Creative	0.571	0.667	0.62

Analysis: A drop in precision is expected (This is no longer an accurate measure of precision as the steps/chains of thought are also listed and included in the metric, whereas only the summary should be extracted, but it will continue to be included for completeness), but an improvement in recall is definitely expected, as each of the models retrieve all the important events/highlights step-by-step.

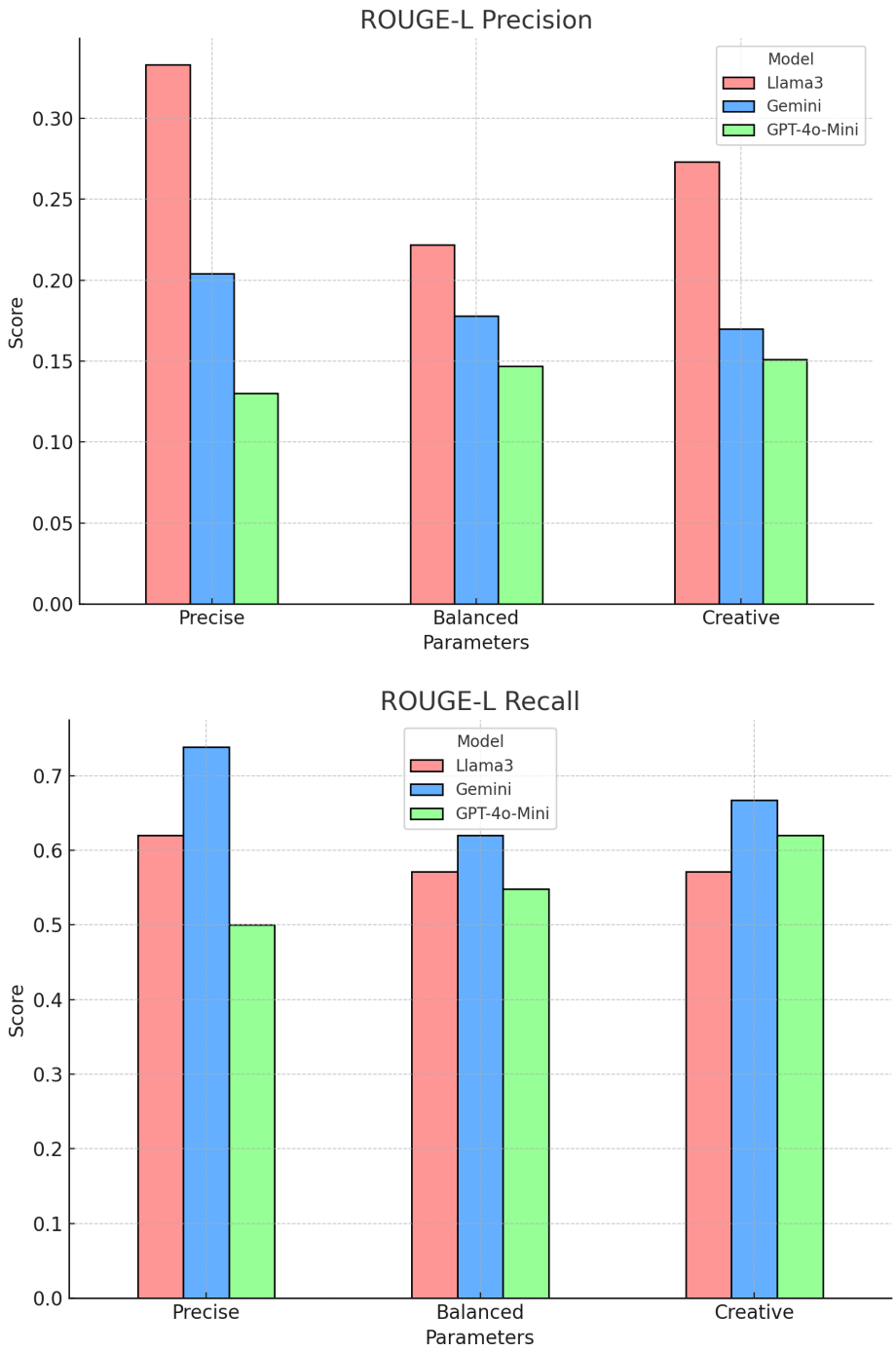


Fig 5: Chain-of-Thoughts (CoT) Text Summarization Metrics

Tree of Thoughts (ToT) Prompting

Here, `max_tokens = 1024`, to allow each “expert” of the model to express its reasoning/strategy.
The ToT prompt for summarization is as follows:

```
    Classify the following news article into one of the categories: World, Sports, Business, Sci/Tech.
    Imagine three different experts are answering this question. All experts will write down 1 step of their thinking, then share it with the group. Then all experts will go on to the next step, etc. If any expert realizes that they're wrong at any point then they leave.
    Article: {article_text}

    Reasoning Paths:
    Path 1:
    1. Identify the main topic of the article.
    2. Determine which category the topic fits into.
    3. Assign the category based on the topic.

    Path 2:
    1. Identify key phrases in the article.
    2. Match key phrases to potential categories.
    3. Assign the category based on the best match.
    Path 3:
    1. Analyze the context and events described in the article.
    2. Compare the context with typical events in each category.
    3. Assign the category based on the closest match.

    Best Path:
```

ROUGE-L Precision

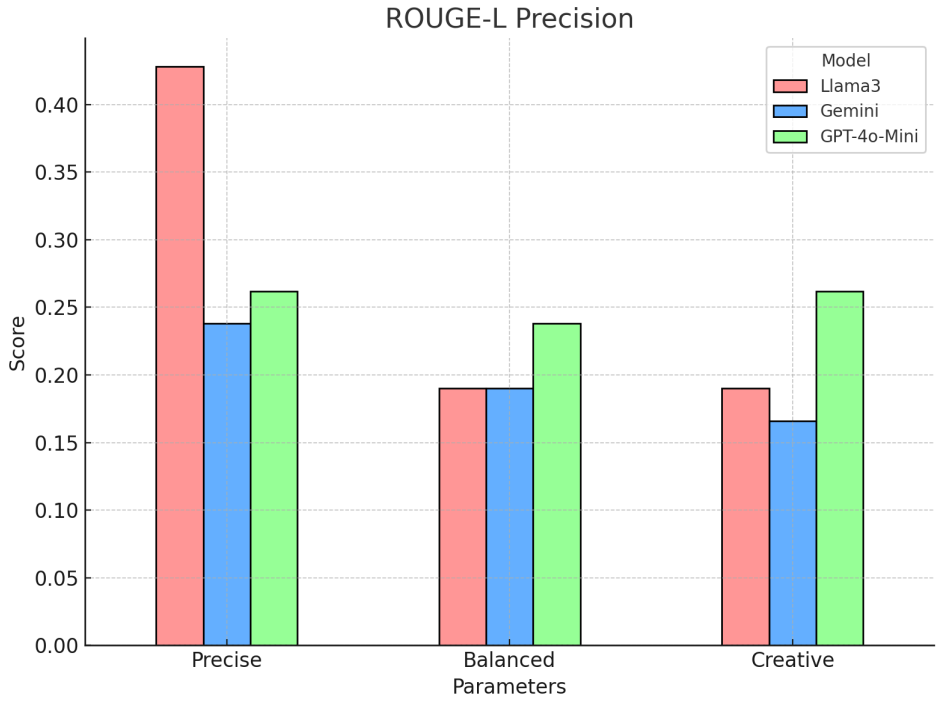
Parameters / Model	Llama3	Gemini	GPT-4o-Mini
Precise	0.428	0.238	0.262
Balanced	0.19	0.19	0.238
Creative	0.19	0.166	0.262

ROUGE-L Recall

Parameters / Model	Llama3	Gemini	GPT-4o-Mini
Precise	0.25	0.151	0.122
Balanced	0.121	0.114	0.112

Creative	0.114	0.111	0.121

Analysis: The metrics are a little more accurate because the summaries were manually extracted. It is worth noting that there is a slight trend reversal where Gemini is producing shorter summaries than Llama3, however it does appear to be at the cost of losing valuable information since both its precision and recall are generally lower than that of Llama3. GPT provides the most verbose summaries as usual but now it also appears that it has gained some precision while trading some recall in the process of doing so. Each of the models seem to have tweaked their summaries after the listing of pros/cons by each “expert” in the thought process.



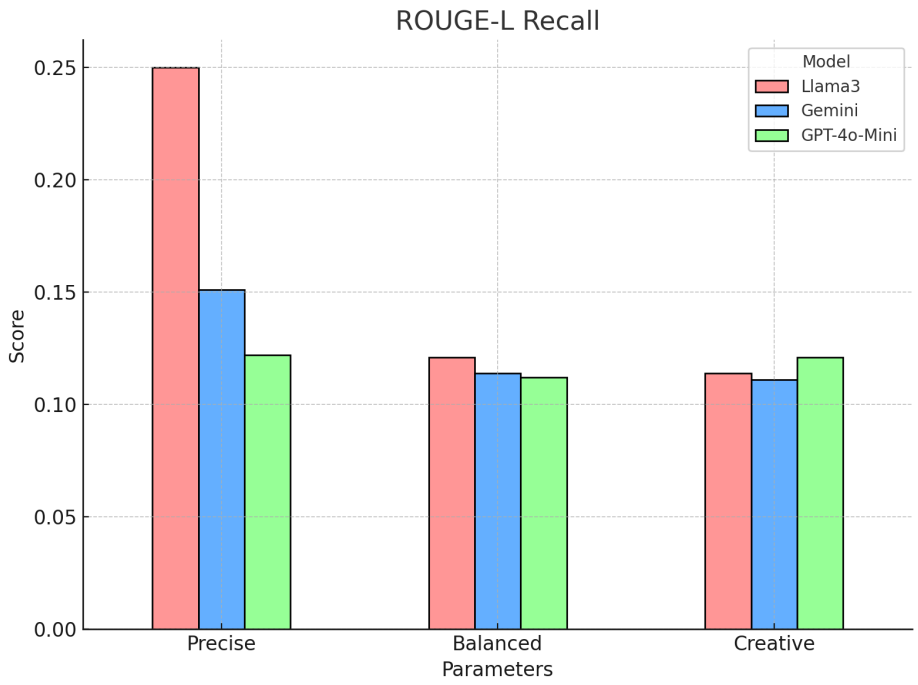


Fig 6: Tree of Thoughts (ToT) Text Summarization Metrics

Conclusion: Llama3 seems to be the best model for summarization with precise and concise summaries. The amount of examples and guidance through examples/thoughts required appear to be minimal for the summarization task, with more guidance “overshooting” and leading the models astray in their summarization. Hence, the One-Shot method of prompting or simply “Let’s think step-by-step” seems to be working the best here.