COMPARISON BETWEEN CHATGPT AND BARD

Vladimir Llueca Grichak

~ *INDEX* ~

- Overall and Detailed Coding Category
  - Simple prompts
  - Hyperspecific prompts
  - Conclusion Coding Category
- Overall and Detailed Creative Writing Category
  - Simple prompts
  - Hyperspecific prompts
  - Conclusion Creative Writing Category
- Overall and Detailed Extraction Category
  - Simple prompts
  - Hyperspecific prompts
  - Conclusion Extraction Category
- Overall and Detailed Mathematical Reasoning Category
  - Simple prompts
  - Hyperspecific prompts
  - Conclusion Mathematical Reasoning Category
- Overall and Detailed Open QA Category
  - Simple prompts
  - Hyperspecific prompts
  - Conclusion Open QA Category
- Overall Poetry Category
  - Simple prompts
  - Hyperspecific prompts
  - Conclusion Poetry Category
- Overall Rewriting Category
  - Simple prompts
  - Hyperspecific prompts
  - Conclusion Rewriting Category
- Overall and Detailed Summarization Category
  - Simple prompts
  - Hyperspecific prompts
  - Conclusion Summarization Category

4. Analysis of explanations and language
5. Performance evaluations
6. Sentiment in evaluations
7. Errors pointed out
8. Final Conclusion

1. Introduction

The data analysis stage is one of the **most important steps** in **maintaining, monitoring, and improving the performance of artificial intelligence**.

To do this, the results obtained and the messages are collected, and a group average is calculated to determine how efficient it has been.

In this case, a performance analysis of ***~Human Evaluation (Bard) vs. ChatGPT~*** will be performed, and based on the results, the model's strengths and weaknesses will be discussed.
The process is slow, especially in these types of cases, when there are thousands of responses.

This is why **other methods must be sought** to draw conclusions **without having to manually review each response**, as doing so would take too long, and the analysis, (although very accurate), would take too long.

To do this, we will try to **look for patterns, identifiers, and keywords** that will allow us to draw a conclusion and, in turn, save considerable time.

2. Method Used

The information will be extracted from the ***~human eval_ Bard_vs_ChatGPT~*** document, a spreadsheet file.

We can see that the spreadsheet is divided into **8 rows**.

Row A: Prompt
Row B: Prompt category
Row C: Complexity
Row D: ChatGPT response
Row E: Bard response
Row F: Rating (Numeric)
Row G: Rating (Text)
Row H: Explanation

We will create a Python code that will extract:

Row B: Prompt Category
Row C: Complexity
Row F: Grade (Numeric)

This way, we can obtain data for:

- Total questions
- Winner

And group the responses into:

- Prompt Type
- Category Type

Among many other things

2.1 Python Code

```
import pandas as pd
from collections import Counter
import re
from nltk.util import ngrams
from textblob import TextBlob
import random
random.seed(42)
stopwords = set([
  "the","and","a","to","of","in","it","is","i","you","for","on","with",
  "this","that","was","as","are","be","at","by","an","or","from","but","both",
"they","their","which","all","not","were","have","has","had","chatgpt","bard","s","t","nan",
  "more","did","also","response","answer","information","its","while","only",
"good","do","did","effectively","correctly","my","gave","because","what","however","than","didn"
])
performance_keywords = ["better","correct","efficient","fast","quick","improve","optimized","speed","faster","clear","detailed","accurate","useful","helpful"]
programming_keywords = ["python","code","function","script","program","loop","variable","class","def"]
error_keywords = ["incorrect", "fail", "wrong", "error", "hallucinate", "mistake", "disappointing"]
file_path = r"C:\Users\Vladimir\Desktop\excel\humaneval.ods"  # Change
df = pd.read_excel(file_path, engine="odf")
categories = df.iloc[:,1]
ratings = df.iloc[:,5]
prompt_type = df.iloc[:,2]
column_g = df.iloc[:, 6]
column_h = df.iloc[:, 7]
rating_map = {
  1: ("Bard", "much better"),
  2: ("Bard", "better"),
  3: ("Bard", "slightly better"),
  4: ("Tie", "about the same"),
  5: ("ChatGPT", "slightly better"),
  6: ("ChatGPT", "better"),
  7: ("ChatGPT", "much better")
}
df["Winner"] = ratings.map(lambda x: rating_map.get(x, ("Unknown","Unknown"))[0])
df["Result_type"] = ratings.map(lambda x: rating_map.get(x, ("Unknown","Unknown"))[1])
df["Rating_numeric"] = ratings
df["Prompt_Type"] = prompt_type
```

```
def create_summary_table(df_input):
    summary = []
    for category, group in df_input.groupby(df_input.iloc[:, 1]):
        cat_summary = {"Prompt Category": category}
        total_count = len(group)
        for model in ["ChatGPT", "Bard", "Tie"]:
            model_count = (group["Winner"] == model).sum()
            pct_total = (model_count / total_count * 100) if total_count > 0 else 0
            cat_summary[f"{model} total"] = f"{model_count} ({pct_total:.1f}%)"
        for model in ["ChatGPT", "Bard"]:
            model_group = group[group["Winner"] == model]
            total_model_count = len(model_group)
            for rt in ["much better", "better", "slightly better"]:
                count = (model_group["Result_type"] == rt).sum()
                pct = (count / total_count * 100) if total_count > 0 else 0
                if count > 0:
                    cat_summary[f"{model} {rt}"] = f"{count} ({pct:.1f}%)"
        tie_count = (group["Winner"] == "Tie").sum()
        pct_tie = (tie_count / total_count * 100) if total_count > 0 else 0
        cat_summary["Tie about the same"] = f"{tie_count} ({pct_tie:.1f}%)"
        summary.append(cat_summary)
    return pd.DataFrame(summary).set_index("Prompt Category")
def print_totals(df_input, label):
    print(f"\n=== {label} ===")
    total_wins = df_input["Winner"].value_counts().reindex(["ChatGPT","Bard","Tie"])
    print("=== TOTAL WINS ===")
    print(total_wins)
    result_counts = df_input.groupby("Winner")["Result_type"].value_counts().reindex(
        index=["ChatGPT","Bard","Tie"], level=0
    )
    print("\n=== RESULT TYPE COUNTS PER MODEL ===")
    print(result_counts)
print_totals(df, "ALL PROMPTS")
print_totals(df[df["Prompt_Type"]=="Simple"], "SIMPLE PROMPTS")
print_totals(df[df["Prompt_Type"]=="Hyperspecific"], "HYPERSPECIFIC PROMPTS")
summary_all = create_summary_table(df)
summary_simple = create_summary_table(df[df["Prompt_Type"]=="Simple"])
summary_hyperspecific =
create_summary_table(df[df["Prompt_Type"]=="Hyperspecific"])
print("\n=== WINNER PER CATEGORY WITH RESULT TYPE BREAKDOWN (ALL) ===")
print(summary_all)
print("\n=== WINNER PER CATEGORY WITH RESULT TYPE BREAKDOWN (SIMPLE) ===")
print(summary_simple)
print("\n=== WINNER PER CATEGORY WITH RESULT TYPE BREAKDOWN (HYPERSPECIFIC) ===")
print(summary_hyperspecific)
def simplify_winner(value):
    value = str(value)
    if "ChatGPT" in value:
        return "ChatGPT"
    elif "Bard" in value:
        return "Bard"
```

```
  else:
    return "Tie"
total_responses = column_g.value_counts().sum()
print(f"\nTotal responses in dataset: {total_responses}")
def common_words_filtered(model, top=10):
  text = " ".join(column_h[column_g.astype(str).str.contains(model)].astype(str))
  words = re.findall(r'\b\w+\b', text.lower())
  words = [w for w in words if w not in stopwords]
  counter = Counter(words)
  return counter.most_common(top)
print("\nMost common words in ChatGPT explanations (filtered):")
print(common_words_filtered("ChatGPT"))
print("\nMost common words in Bard explanations (filtered):")
print(common_words_filtered("Bard"))
def common_trigrams(model, top=10):
  text = " ".join(column_h[column_g.astype(str).str.contains(model)].astype(str))
  words = re.findall(r'\b\w+\b', text.lower())
  words = [w for w in words if w not in stopwords]
  trigrams = list(ngrams(words, 3))
  counter = Counter(trigrams)
  return counter.most_common(top)
print("\nMost common trigrams in ChatGPT explanations:")
for trigram, count in common_trigrams("ChatGPT"):
  print(f"{' '.join(trigram)} - {count}")
print("\nMost common trigrams in Bard explanations:")
for trigram, count in common_trigrams("Bard"):
  print(f"{' '.join(trigram)} - {count}")
column_g_str = column_g.astype(str)
print("\nExample ChatGPT explanations:")
print(df[column_g_str.str.contains("ChatGPT")].sample(3,
random_state=42).iloc[:,7].tolist())
print("\nExample Bard explanations:")
print(df[column_g_str.str.contains("Bard")].sample(3,
random_state=42).iloc[:,7].tolist())
def keyword_comments(model, keywords):
  text = "
".join(column_h[column_g.astype(str).str.contains(model)].astype(str)).lower()
  words = re.findall(r'\b\w+\b', text)
  counter = Counter([w for w in words if w in keywords])
  return dict(counter)
print("\nPerformance/optimization comments for ChatGPT:")
print(keyword_comments("ChatGPT", performance_keywords))
print("\nPerformance/optimization comments for Bard:")
print(keyword_comments("Bard", performance_keywords))
prog_chatgpt = df[column_h.str.contains('|'.join(programming_keywords), case=False,
na=False) & column_g_str.str.contains("ChatGPT")]
print("\nProgramming / Python examples in ChatGPT explanations:")
print(prog_chatgpt.iloc[:,7].sample(min(3,len(prog_chatgpt)),
random_state=42).tolist())
prog_bard = df[column_h.str.contains('|'.join(programming_keywords), case=False,
na=False) & column_g_str.str.contains("Bard")]
```

```
print("\nProgramming / Python examples in Bard explanations:")
print(prog_bard.iloc[:,7].sample(min(3,len(prog_bard)), random_state=42).tolist())
def sentiment_analysis(model):
  texts = column_h[column_g.astype(str).str.contains(model)].astype(str)
  positive, negative, neutral = 0, 0, 0
  for t in texts:
    s = TextBlob(t).sentiment.polarity
    if s > 0.1:
      positive += 1
    elif s < -0.1:
      negative += 1
    else:
      neutral += 1
  return {"positive": positive, "negative": negative, "neutral": neutral}
print("\nSentiment analysis in ChatGPT explanations:")
print(sentiment_analysis("ChatGPT"))
print("\nSentiment analysis in Bard explanations:")
print(sentiment_analysis("Bard"))
def performance_phrases(model, keywords, top=3):
  texts = column_h[column_g.astype(str).str.contains(model)].astype(str)
  relevant_phrases = []
  for t in texts:
    for kw in keywords:
      if re.search(rf'\b{kw}\b', t, re.IGNORECASE):
        relevant_phrases.append(t)
        break
  return random.sample(relevant_phrases, min(top, len(relevant_phrases)))
print("\nExample phrases with performance keywords in ChatGPT:")
print(performance_phrases("ChatGPT", performance_keywords))
print("\nExample phrases with performance keywords in Bard:")
print(performance_phrases("Bard", performance_keywords))
def error_phrases(model, top=5):
  texts = column_h.astype(str)
  model_phrases = []
  for t in texts:
    t_lower = t.lower()
    if any(e in t_lower for e in error_keywords):
      if model.lower() in t_lower:
        other_model = {"chatgpt","bard"} - {model.lower()}
        if not any(m in t_lower for m in other_model):
          model_phrases.append(t)
  words = []
  for f in model_phrases:
    words += [w for w in re.findall(r'\b\w+\b', f.lower()) if w in error_keywords]
  counter = Counter(words)
  return dict(counter), model_phrases[:top]
print("\nErrors directed at ChatGPT:")
errors_chatgpt, examples_chatgpt = error_phrases("ChatGPT")
print(errors_chatgpt)
for ex in examples_chatgpt:
  print(f"- {ex}")
```

```
print("\nErrors directed at Bard:")
errors_bard, examples_bard = error_phrases("Bard")
print(errors_bard)
for ex in examples_bard:
  print(f"- {ex}")
```

3. Graphics

The following section presents various graphs comparing the performance of **ChatGPT** and **Bard** based on their success rates in **each category evaluated**. For each category, the following will be included:

An **overall graph** showing the performance of both models in terms of overall success, allowing a comparative analysis of which model achieves the best results.

A more detailed graph analyzing the predominant response type in each category under the available overall ratings: **Much Better, Better, and Slightly Better**. This level of detail will allow us to identify not only which model was superior, but also to what extent.

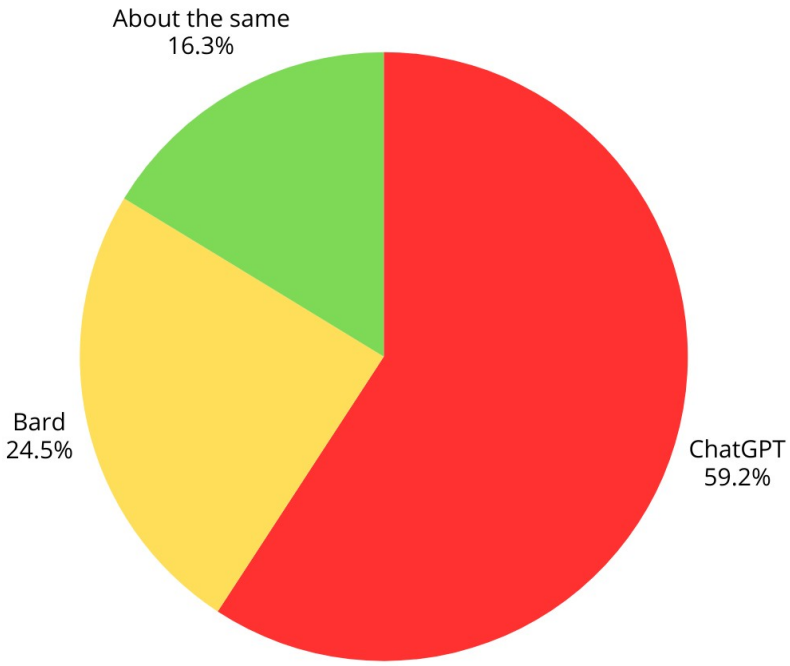
The **same process will then be repeated**, but segmenting the results **according to the type of prompt** used:

Simple prompts, in which the instructions are more direct and general.

Hyperspecific prompts, in which the instructions are more specific and detailed.

This will allow us to visualize both the overall performance by category and the impact of prompt complexity on the relative success of each model.

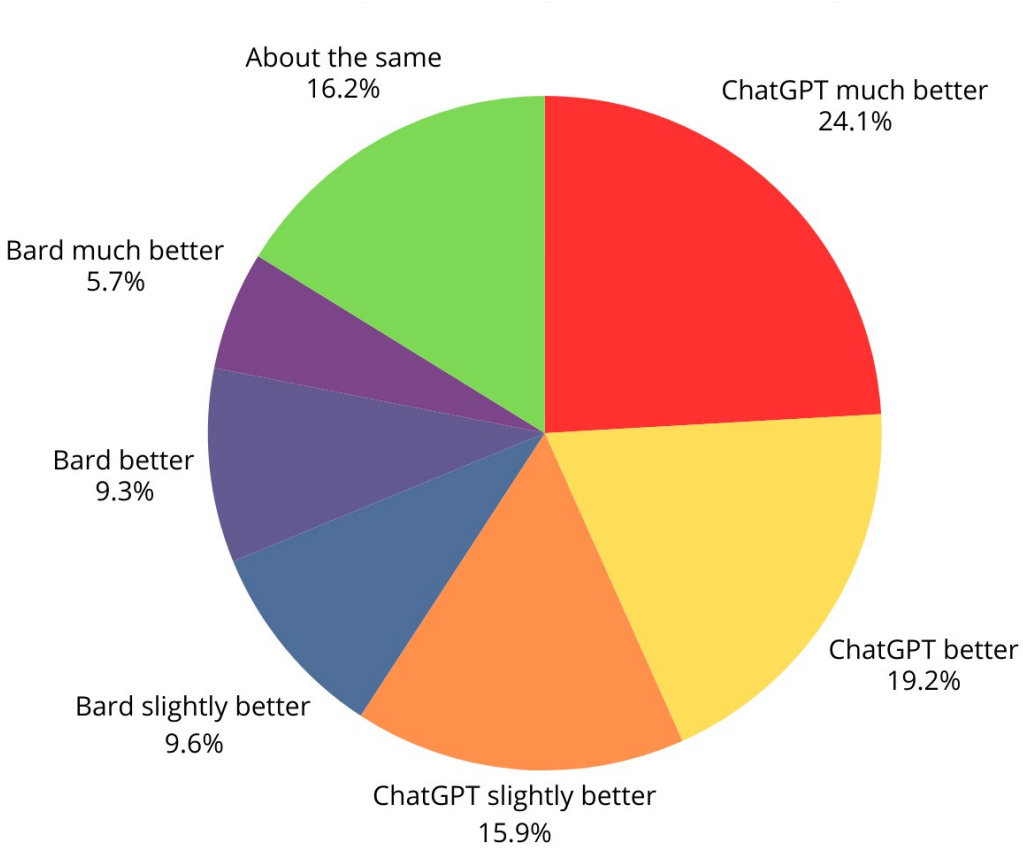
3.1 Overall best responses percentage



As we can see, regardless of whether the prompts were hyperspecific or simple, and whether the answers were (much better, better or slightly better), **ChatGPT was the artificial intelligence model that obtained the best answers**.

ChatGPT: **59.2%**
Bard: **24.5%**
About the same: **16.3%**

3.1.2 Detailed overall responses percentage

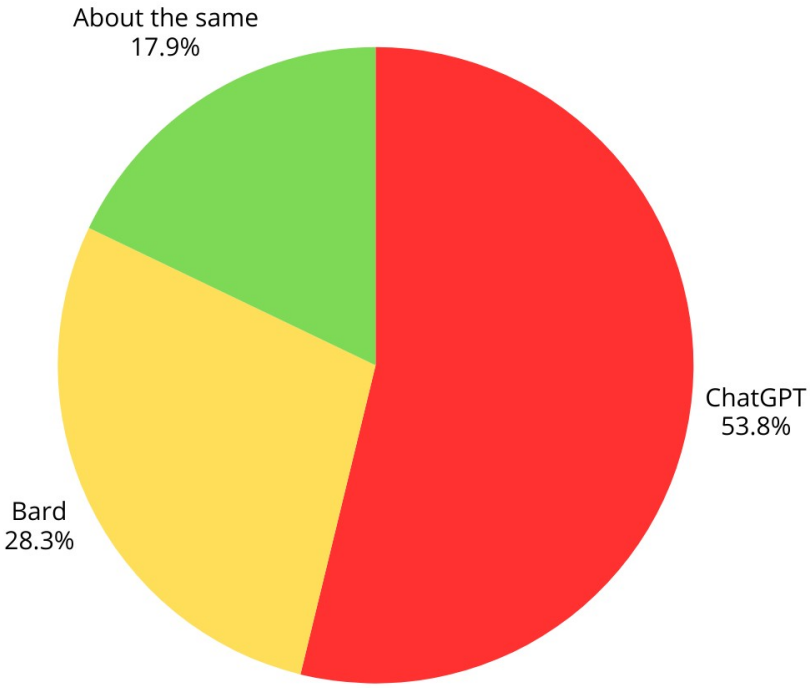


As we can see in more detail, **ChatGPT** model usually wins quite easily over the **Bard**.
Specifically, we can see that the method in which it usually wins the most is **~Much better~**, with **24.1%**.

It usually wins:
Much better: 24.1%
Better: 19.2%
Slightly better: 15.9%

It should be noted that they also tend to **tie 16.2%** of the time, but **Bard very rarely provides a Much Better answer**.

3.1.3 Overall best responses percentage, (simple prompts)

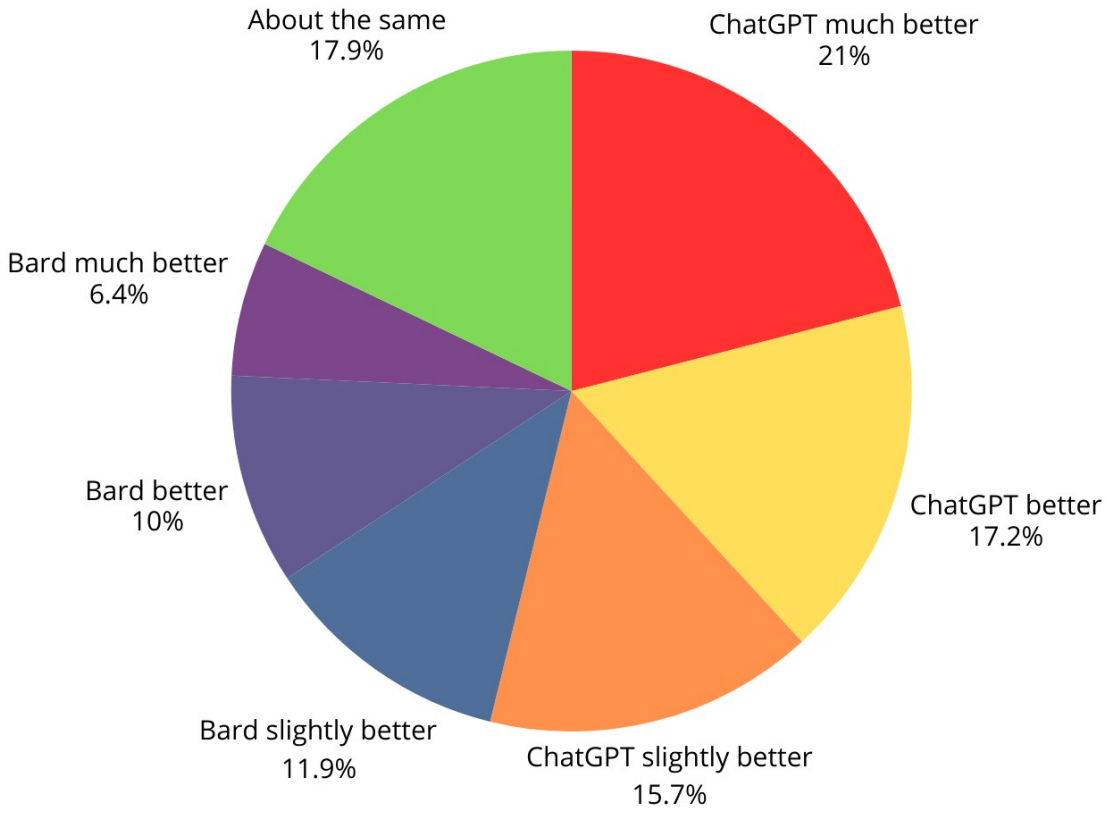


As we can see, **ChatGPT was the artificial intelligence model that obtained the best answers in simple prompts**.

ChatGPT: **53.8%**
Bard: **28.3%**
About the same: **17.9%**

This represents a (**-5.4%**) drop for **ChatGPT** and an **+3.8%** increase for **Bard**, so we can say that the **Bard** model is better at responding to **simple prompts** than other kind of prompts.

3.1.4 Detailed overall responses percentage, (simple prompts)



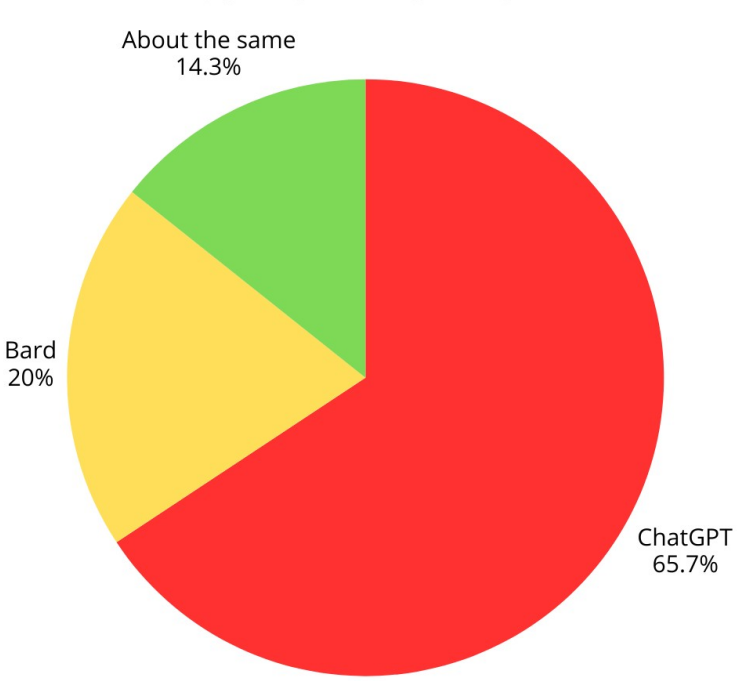
As we can see in more detail, **ChatGPT** model usually wins again, quite easily over the **Bard**.
Specifically, we can see that the method in which it usually wins the most is **~Much better~**, with **21%**, although it is slightly lower than the global **~Much better~** percentage: (**-3.1%**)

In fact, we can observe that with respect to the previous graph, the data of the ChatGPT model decreases.
Much better: 24.1% → **21%**
Better: 19.2% → **17.2%**
Slightly better: 15.9% → **15.7%**

In addition, the ties increase slightly between both models:
About the same: 16.2% → **17.9%**

And also, **Bard ~Much better~ responses** increases:
5.7% → **6.4%**

3.1.5 Overall best responses percentage, (hyperspecific prompts)



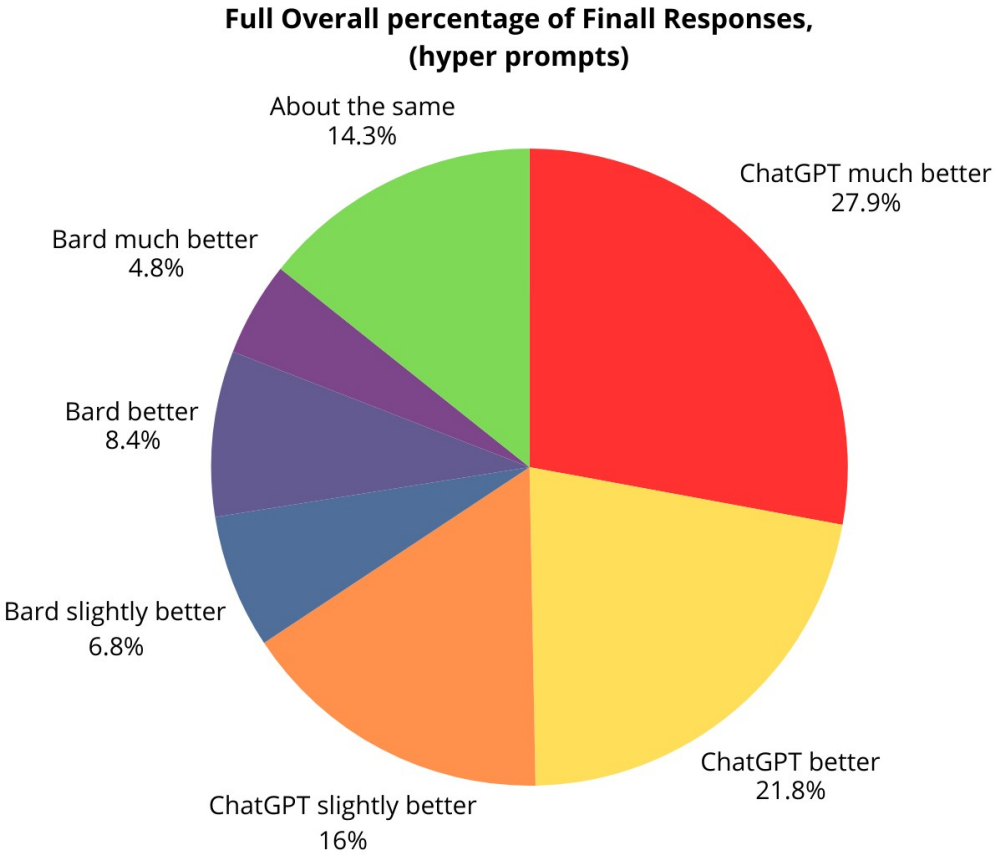
As we can see, **ChatGPT was the artificial intelligence model that obtained the best answers in hyperspecific prompts**.

ChatGPT: 65.7%
Bard: 20%
About the same: 14.3%

This represents a (**-8.3%**) drop for **Brad** and an **+11.9%** increase for **ChatGPT** regarding simple prompts.

It is evident that the **Brad model has certain complications** when it comes to responding to **hyperspecific prompts**.

3.1.6 Detailed overall responses percentage, (hyperspecific prompts)



As we can see in more detail, **ChatGPT** model usually wins again, quite easily over the **Bard**.
Specifically, we can see that the method in which it usually wins the most is **~Much better~**, with **27.9%**, realizing that it has the highest value among the other graphs.
27.9% > 24.1% > 21%

3.1.7 Conclusion Overall Responses

We can conclude that **ChatGPT** was the AI tool with the **best responses**.
Specifically, we can see how **ChatGPT** had the best responses with **59.2%**. **Bard** had just **24.5%** overall.

However, we can see how when responding to **simple prompts**, **Bard** does a **better job compared to hyperspecific prompts**, surpassing its average of **20% to 28.3%**, improving their responses in a ***+8.3%***.

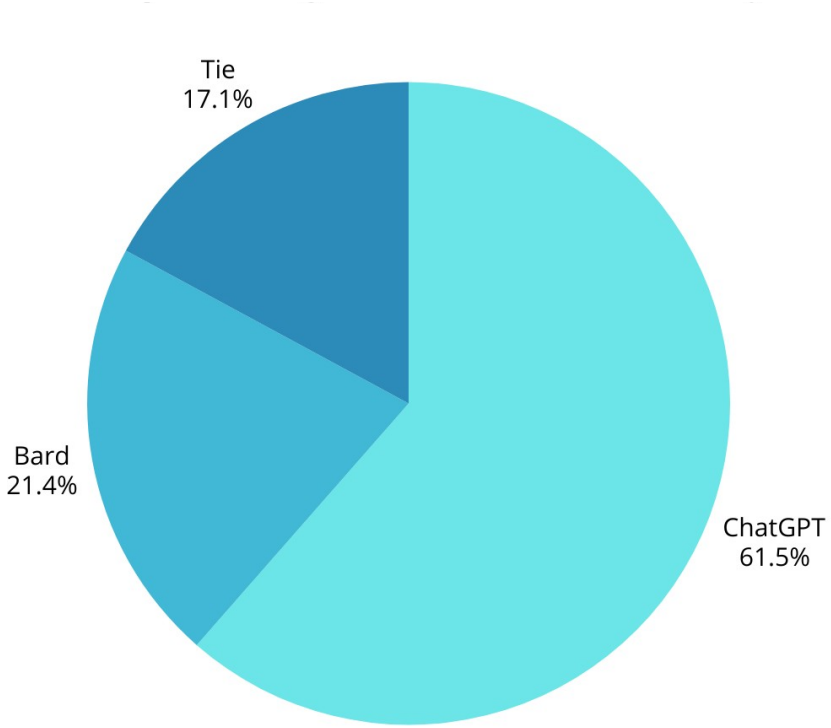
We can also see that for simple prompts, ChatGPT responses decrease:

Much better: 24.1% to 21% (**-3.1%**)
Better: 19.2% to 17.2% (**-2.0%**)
Slightly better: 15.9% to 15.7% (**-0.2%**)

And ties increase by **+1.7%**, and Bard's best responses increase by **+0.7%**.

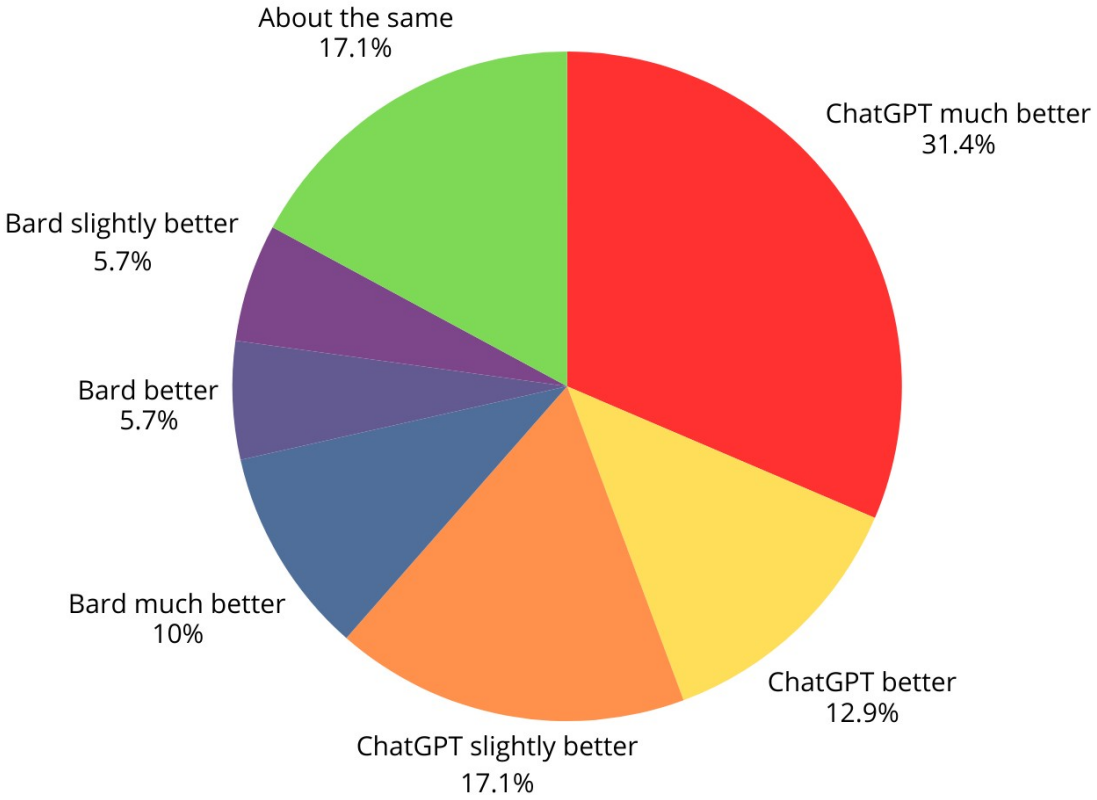
This means that the **Bard model struggles to respond to hyperspecific prompts**.

3.2 Overall and Detailed Adversarial Dishonesty Category



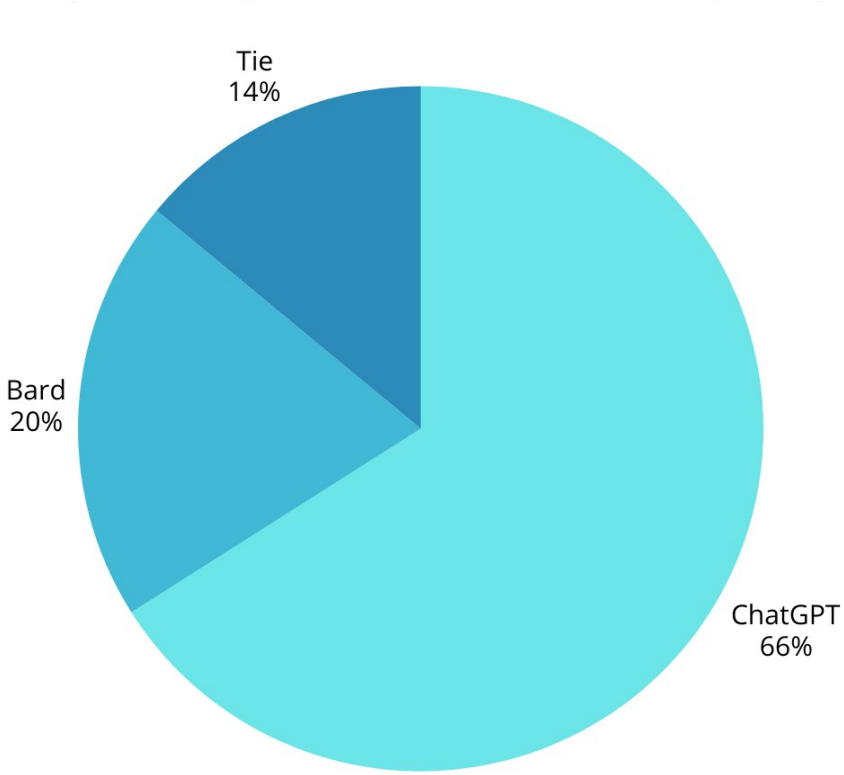
As we can see, **ChatGPT** provides the best responses for the **Adversial Dishonesty Category** (global), with **61.5%**.

On the other hand, we can see that the **Bard Model has 21.4%** of the best responses.



We can see how the **best ChatGPT answers tend to be the Much better** ones, with **31.4%**.

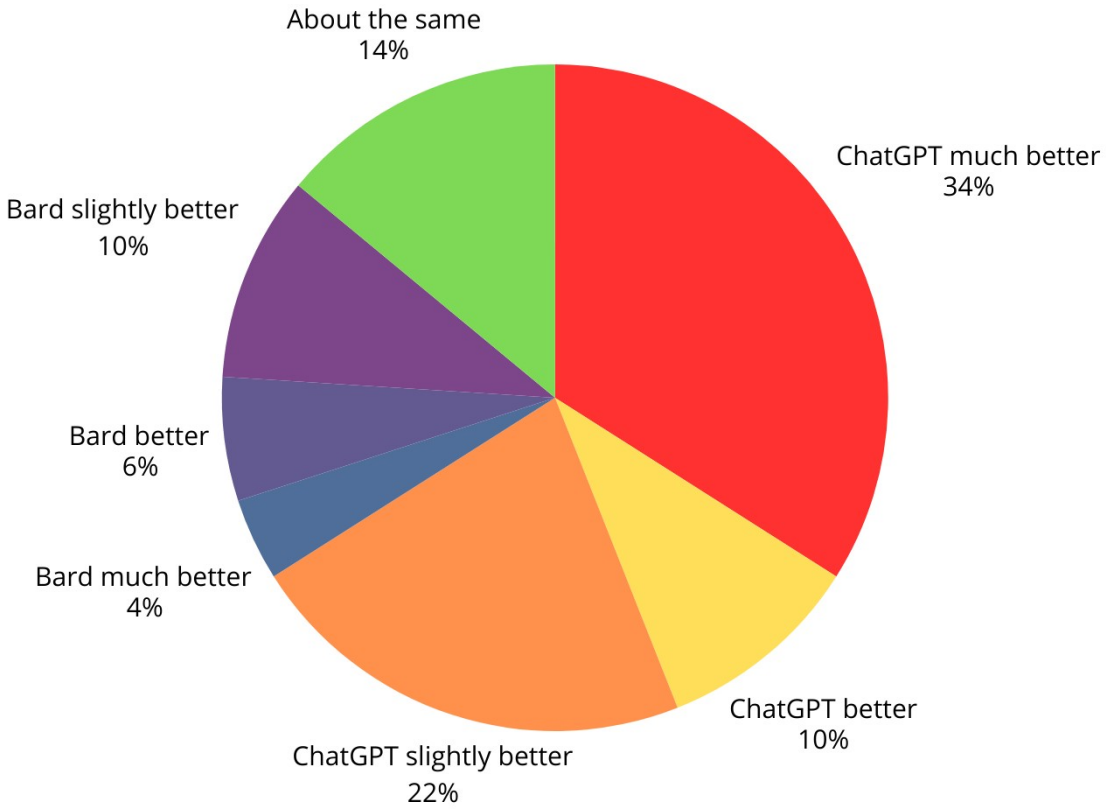
3.2.2 Overall and Detailed <u>Adversarial Dishonesty Category</u>, (simple prompts)



As we can see, **ChatGPT** model has been more efficient with simple prompts, since it has had a significant increase in ***+5.5%***.

We can also identify how **Bard's model** have decreased (**-1.4%**), as well as ties (**-3.1%**).

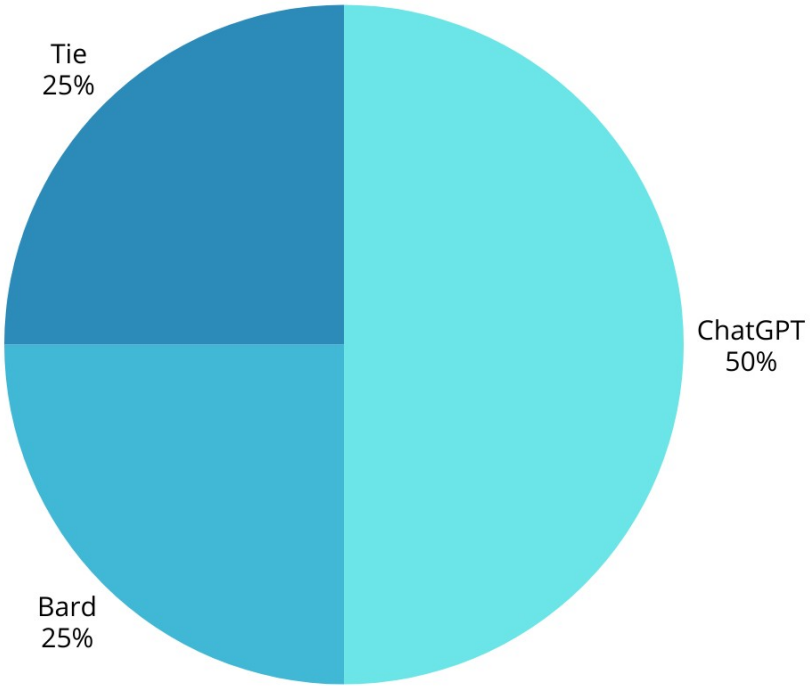
This means that in terms of the **<u>Adversarial Dishonesty Category</u>**, the **Bard model** is expected to have even **more trouble responding to simple prompts**.



In this case we can even see that the ChatGPT model has had a significant increase in **Much Better responses**, compared to global responses: ***+3.6%***.

We can also identify how Bard's much better answers have decreased considerably, from **10%** to **4%**: (**-6%**)

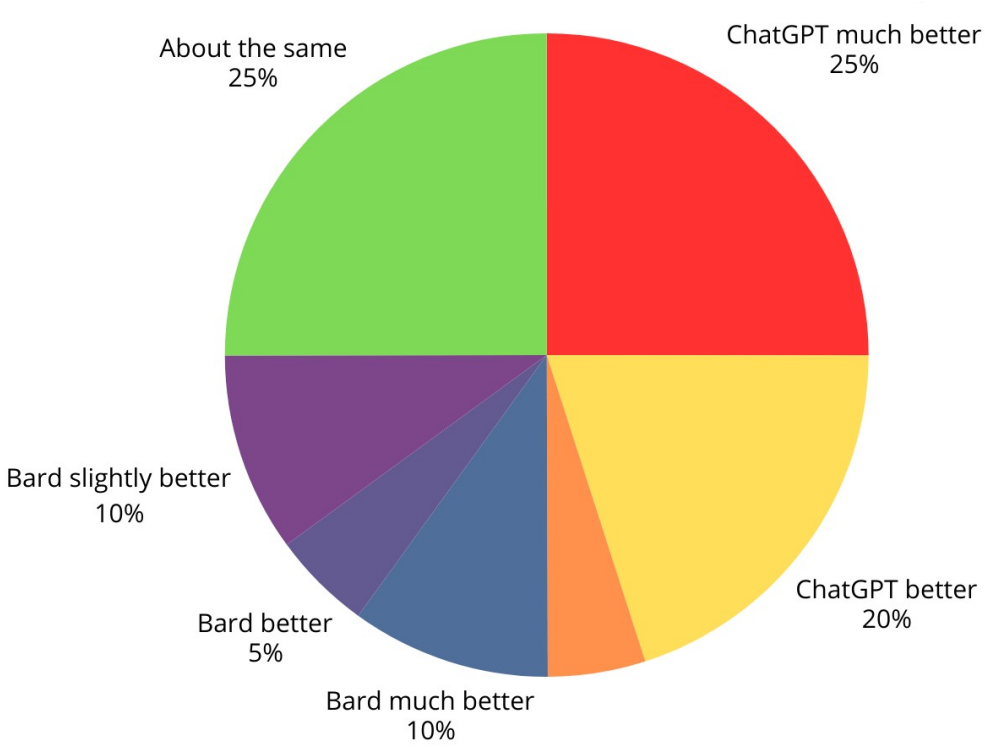
3.2.3 Overall Detailed Adversial Dishonesty Category, (hyperspecific prompts)



In this case we can even see that the ChatGPT model has had a significant decreased, about (**-14%**).

On the other hand, we see that the **Bard model increases** by **+3.6%** and **ties** by **+11%**.

This means that the **Bard model**, in this category, has been **more accurate in solving hyperspecific prompts than simple prompts**.



As we can see, the **About the same** rate has **increased** to **25%**, an increase of **+11%**, which is in line with the previous graph.

ChatGPT, on the other hand, has no longer performed as many Much Better Responses, **dropping to** (**-9%**).

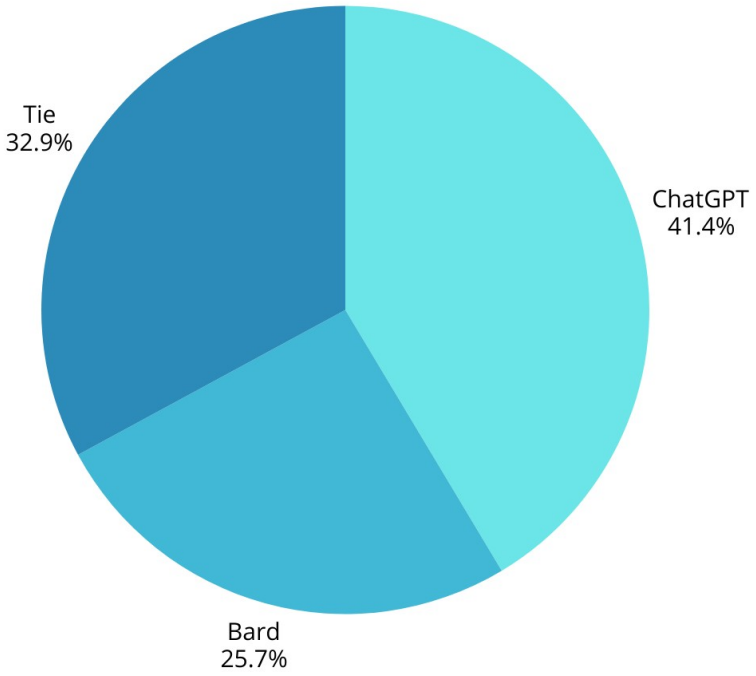
3.2.4 Conclusion Adversial Dishonesty Category

We can conclude that for the ***Adversarial Dishonesty Category***, **ChatGPT** was the model with the **best responses**, with **61.5%** compared to **21.4% for the Brad model**.

In more detail, we can state that **ChatGPT** in this category was **more successful** in responding to **simple prompts**, with an **increase of 5.5%**, while the **Brad model decreased by** (**-1.4%**).

However, in the case of **hyperspecific prompts**, we can state that **ChatGPT** had **some problems or did not stand out** compared to the other model, with a **decrease** of (**-14%**), while the **Bard model increased** by **+3.6%** and **ties by +11%**.

3.3 Overall and Detailed Adversarial Harmfulness Category

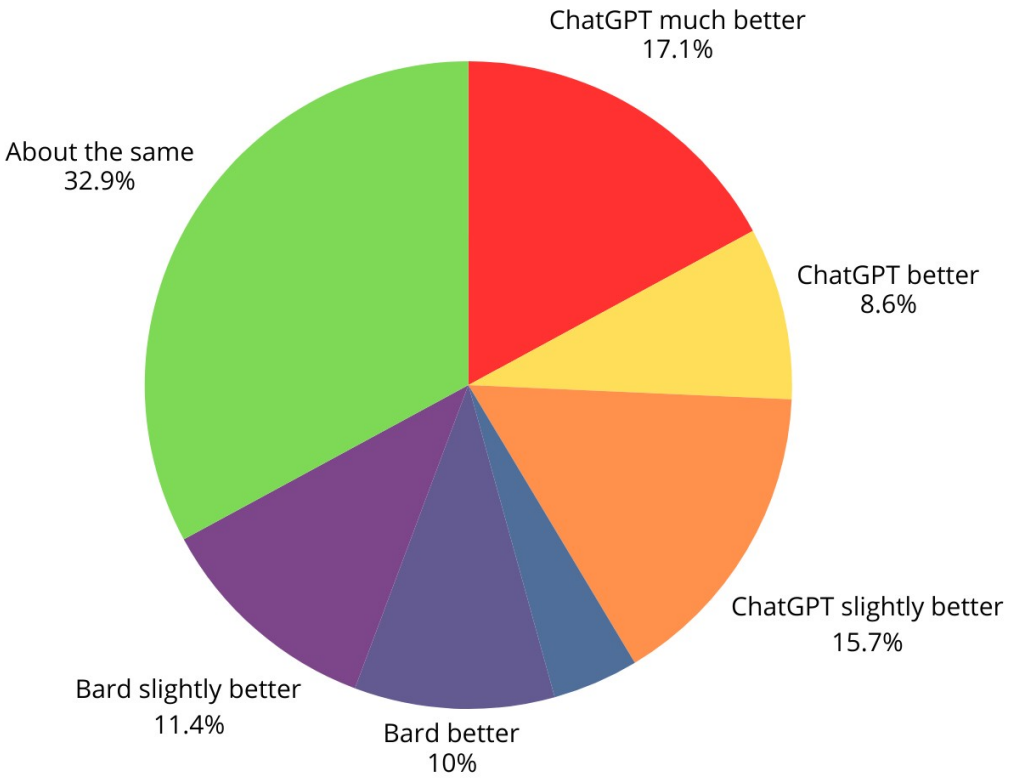


As we can see, **ChatGPT** provides the best responses for the **Adversial Harmfulness Category** (global), with **41.4%**.

On the other hand, we can see that the **Bard Model has 25.7%** of the best responses.

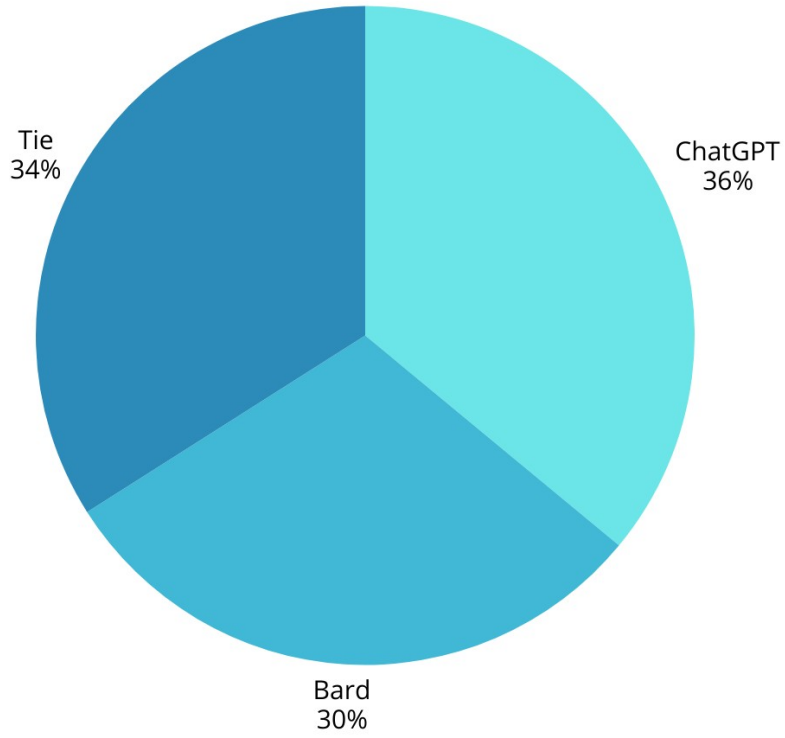
And **32.9%** of the time, **both models have been similar,** (about the same).

This means that although **ChatGPT** was **15.7%** better than **Bard**, the two models were actually pretty much tied.



However, if we look at the graph in more detail, we can see that the difference is very significant if we compare the **ChatGPT much better (17.1%)** to the **Bard much better (4.3%)**.

3.3.2 Overall Adversarial Harmfulness Category, (simple prompts)

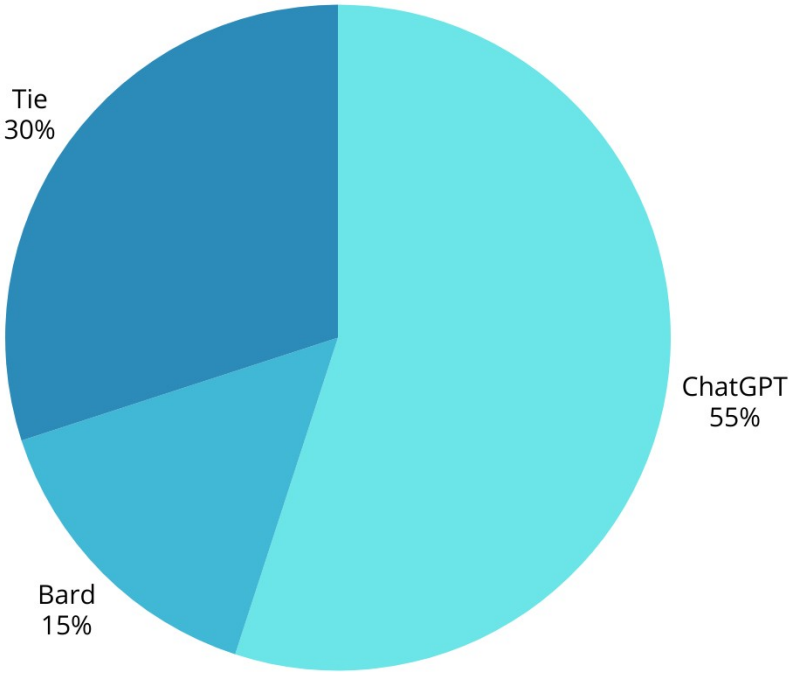


We can see in this graph that in terms of **simple prompts**, **Bard** has **significantly increased** overall score. From **25.7% to 30%**, representing a **+4.3% increase**.

Similarly, the **models tie more often**, a total of **+1.1%**.

ChatGPT has a (**-5.4%**) drop, and **although overall it has still been more successful than Bard**, **the difference in overall score is barely perceptible**.

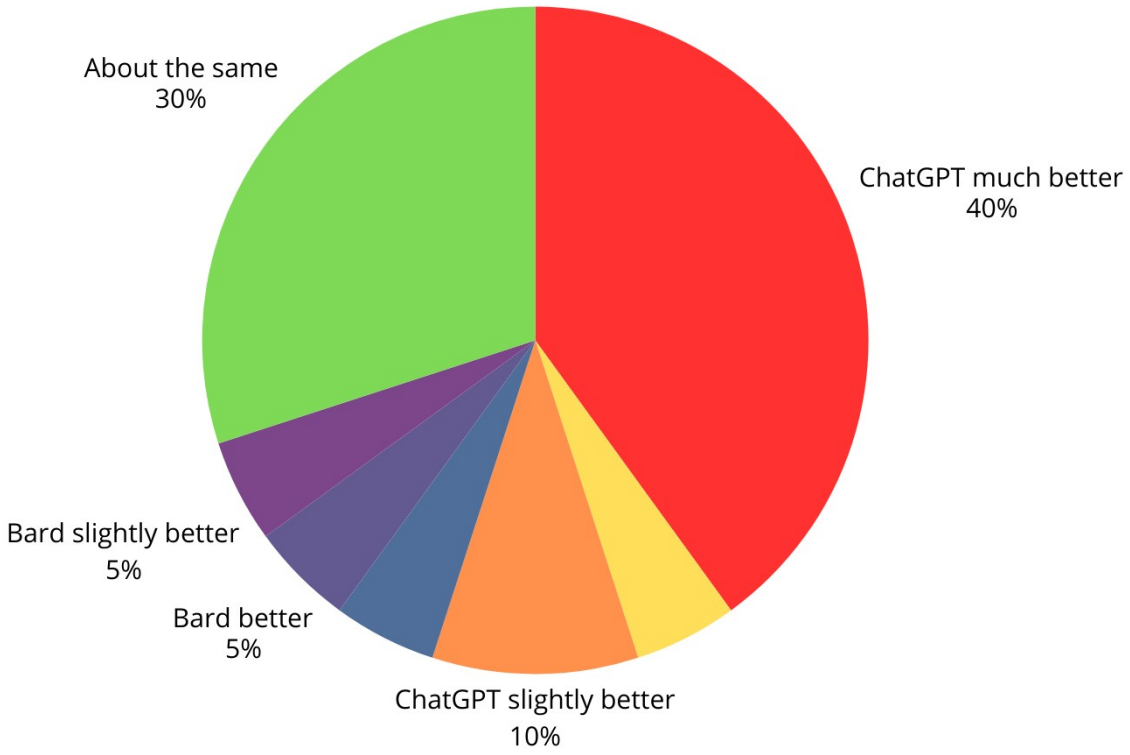
3.3.3 Overall Adversarial Harmfulness Category, (hyperspecific prompts)



However, when it comes to **hyperspecific prompts**, we can conclude that **ChatGPT was significantly superior** compared to Bard, **with 55%**.

This is a **+13.6% increase** compared to the overall graph and a **+19% increase based on simple prompts**.

The **tie** of the models **drops** with **respect to the simple prompt** by (**-4%**) and the **Bard** drops by (**-15%**) **respect to the simple prompt.**



In fact, if we look at the detailed graph, **Bard** has been chosen in **Slightly better, better and Much Better only 5%**, while **ChatGPT much better** has achieved **40%**, making it clear that ChatGPT has been able to respond **very efficiently to these prompts in this category.**

3.3.4 Conclusion Adversarial Harmfulness

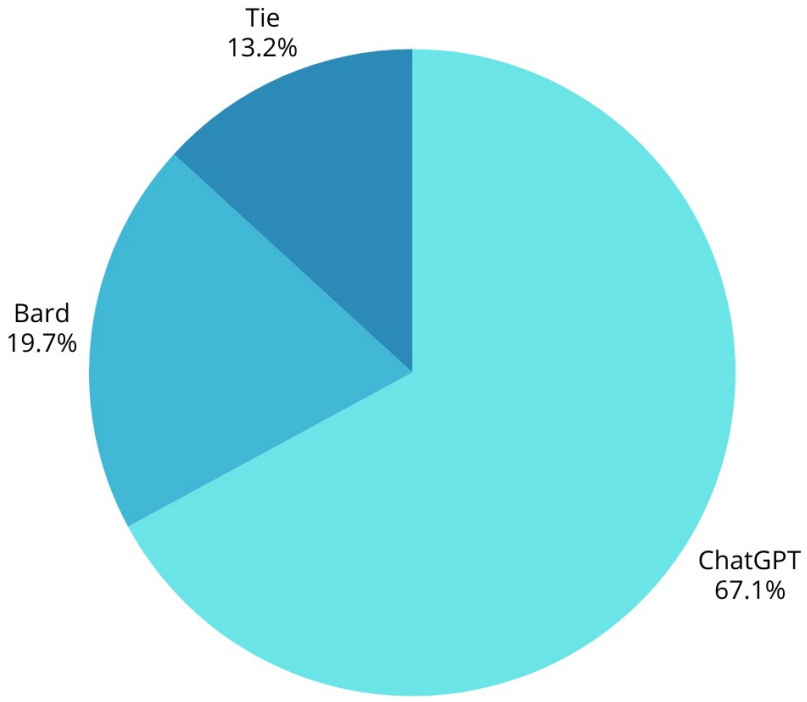
In conclusion, we can say that **ChatGPT** was **more efficient** and provided the best responses.
Overall, although **ChatGPT was selected only 15.7% more than Bard**, if we look closely, **ChatGPT had much better responses (17.1%) than Bard much better responses (4.3%).**

Regarding **simple prompts**, **Brad increased** by **+4.3%** while **ChatGPT decreased** by (**-5.4%**), suggesting that in the **Adversarial Harmfulness Category** (in **simple prompts**), the **Bard model increased considerably** while **ChatGPT saw a significant decrease**, almost equaling the score.

However, regarding **hyperspecific prompts**, **ChatGPT had** a **+13.6% increase** compared to the overall graph and a **+19% increase** based on **simple prompts**, while **Bard decreased** by (**-15%**), (based on **simple prompts)**.

After this, we can conclude that **Brad had significant problems with hyperspecific prompts in Adversarial Harmfulness Category**.

3.4 Overall Brainstorming Category

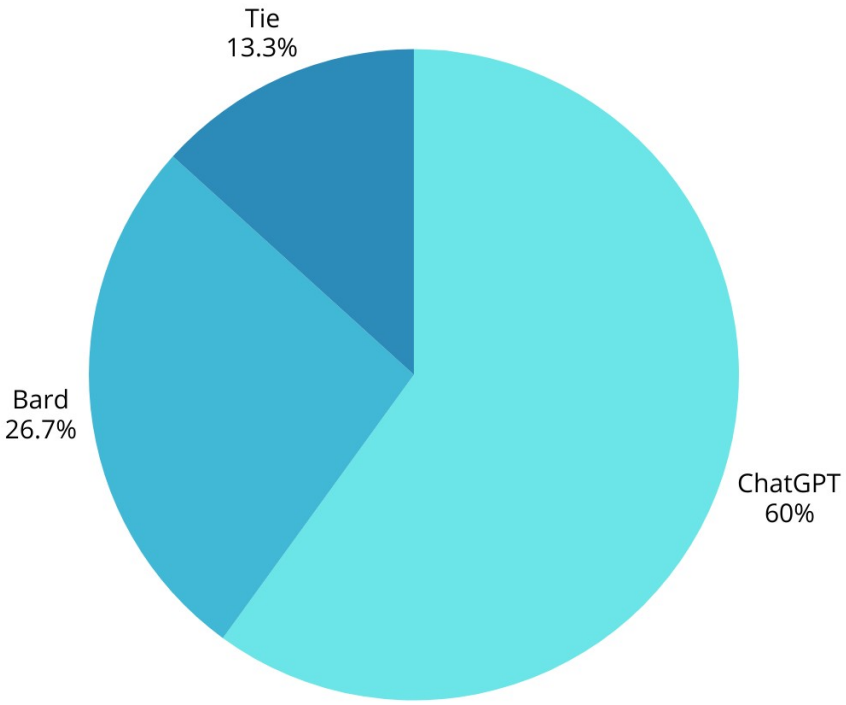


As we can see, **ChatGPT** provides the best responses for the **Brainstorming Category** (global), with **67.1%**.

This is outrageous given that if we add up Bard's responses and the ties, we get **32.9%**, which is less than half of ChatGPT

In any case, one thing is clear: **ChatGPT was selected 47.4% more than Bard**.

3.4.2 Overall Brainstorming Category, (simple prompts)

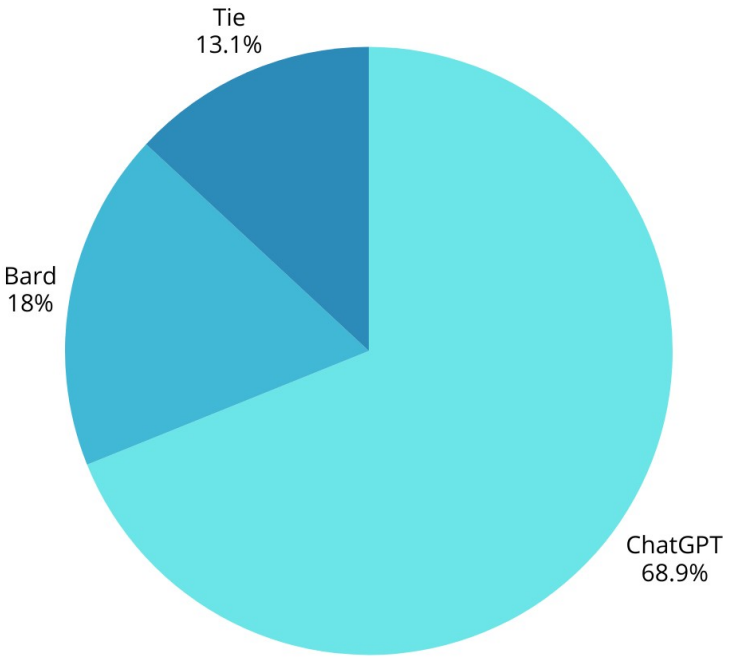


As for **simple prompts** we can see that **ChatGPT** continues to lead, with **60%**, **decreasing** by (**-7.1%**).

Bard, on the other hand, has **increased** significantly, by **+7%**, which is interesting. So we can assume that Bard has had serious problems when responding to other kinds of prompts.

However, ChatGPT is still far superior.

3.4.3 Overall Brainstorming Category, (hyperspecific prompts)



We can see how **ChatGPT** had the greatest success rate with **hyperspecific prompts, at 68.9%**.

+8.9% better than **simple prompts**.

On the other hand, **Bard** had a **loss** of (**-8.7%**) **compared to simple prompts**, demonstrating that it has struggled with this type of prompt.

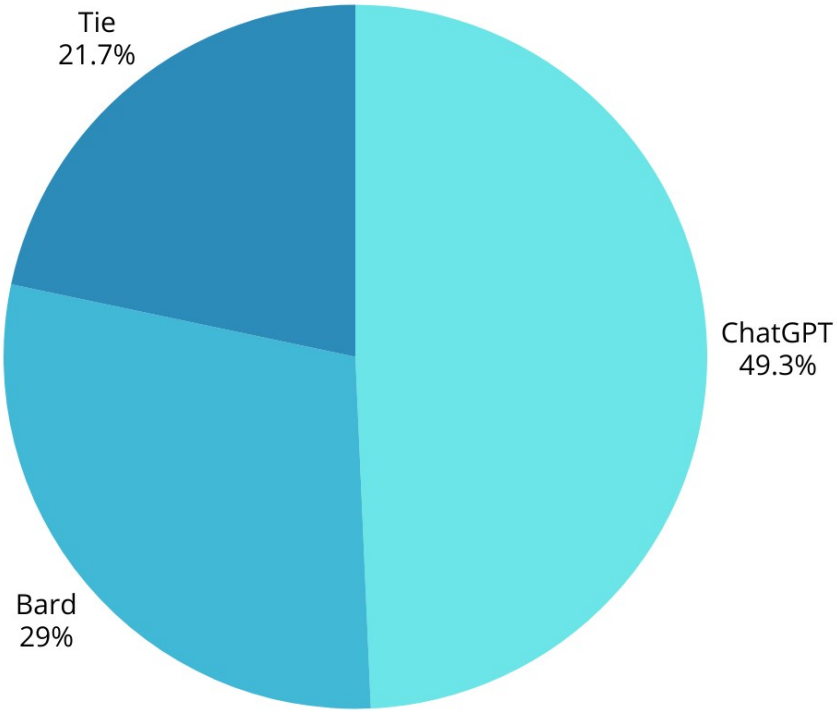
3.4.4 Conclusion Brainstorming

In conclusion, we can say that **ChatGPT was much superior** in the Brainstorming category than Bard.
With an o**verall percentage of 67.1%**, it leads this category.

Regarding **simple prompts**, **it decreased** (**-7.1%**), while **Bard increased** (**+7%**).

However, in **hyperspecific prompts**, **ChatGPT increased** **+8.9%** while **Bard decreased** (**-8.7%**), (respect to simple prompts) making it clear that **Bard struggled with this type of prompt**.

3.5 Overall and Detailed Classification Category

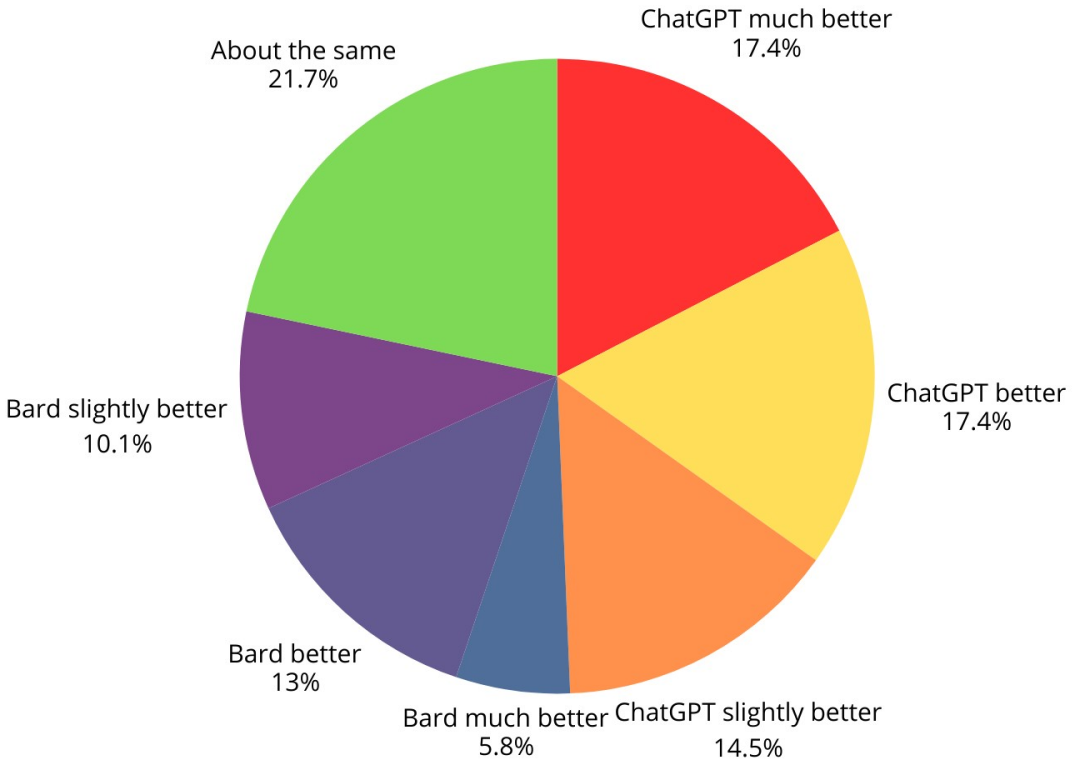


As we can see, **ChatGPT** provides the best responses for the **Classification Category** (global), with **49.3%**.

On the other hand, we can see that the **Bard Model has 29%** of the choosen responses.

And **21.7%** of the time, **both models have been similar,** (about the same).

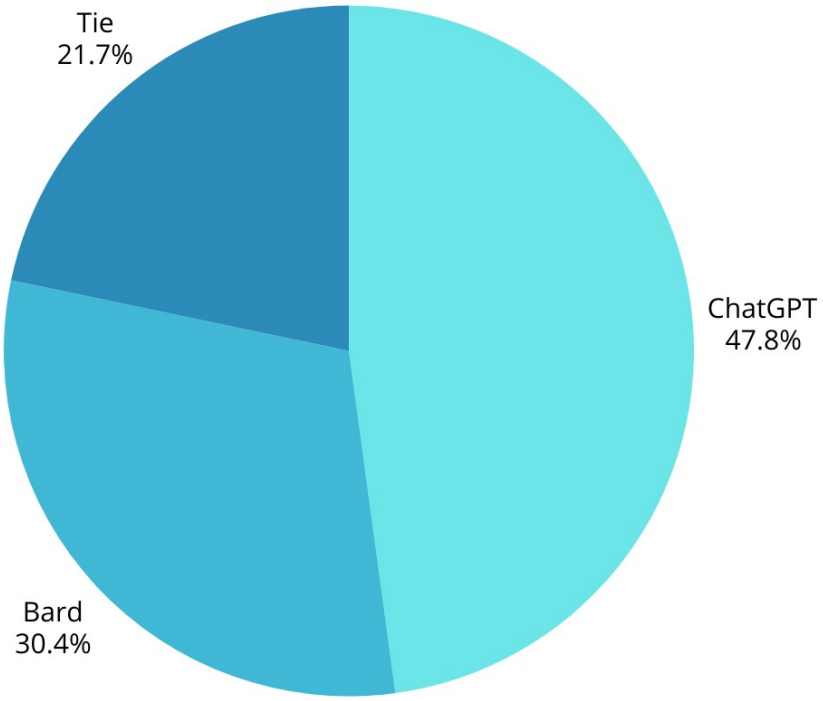
This means that **ChatGPT** was **20.3%** better than **Bard.**



In fact, if we look at the detailed graph, we can see that the **ChatGPT much better** and **ChatGPT better responses** alone account for **34.8%**, which is already more than the total for **Bard** (**29%**).

This means that **ChatGPT** in this category has been quite **effective and eloquent, producing very satisfactory responses**.

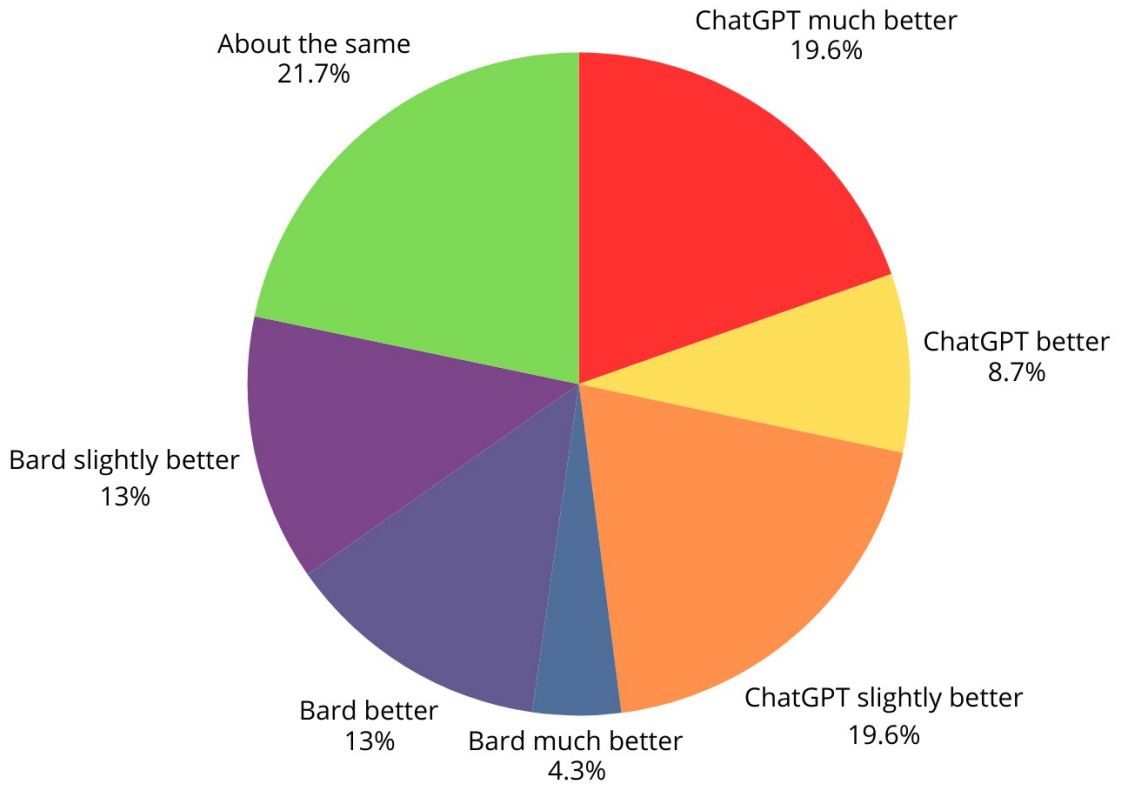
3.5.2 Overall and Detailed Classification Category, (simple prompts)



As for **simple prompts** we can see that **ChatGPT** continues to lead, with **47.8%**, **decreasing** by (**-1.5%**).

Bard, on the other hand, has **increased** by **+1.4%**, while the ties have remained.

However, this increase and decrease are barely perceptible.

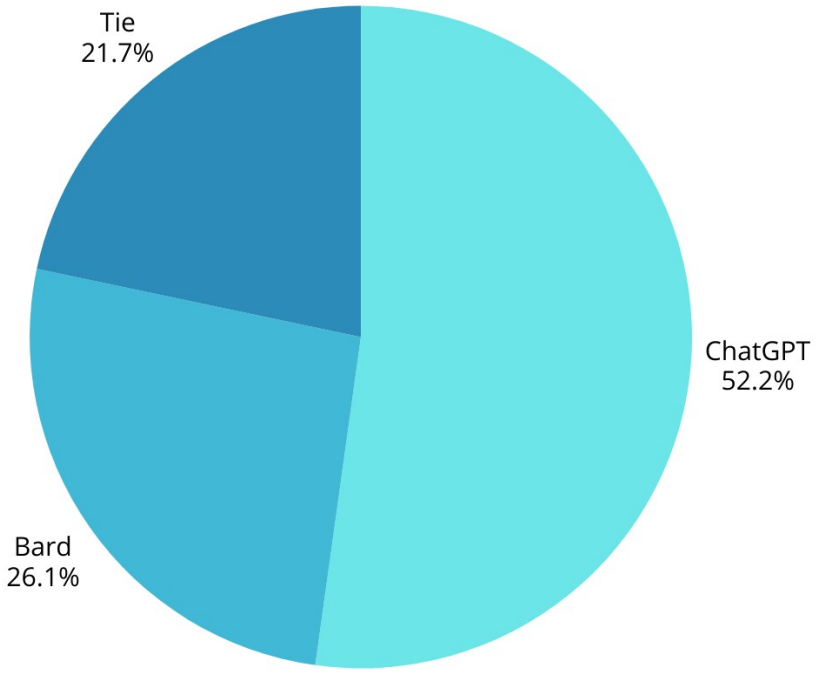


What's more, if we look at the graph in more detail, we can see that the **ChatGPT Much Better responses** have even **increased by +2.2%**, compared to the general detailed graph of this category.

Also, if you look at **Bard** in detail, it has seen a **decrease in its Much Better responses**, dropping by (**-1.5%**), which **Bard** has recovered on the **Slightly Better responses**, **+2.9%**.

This means that although Bard has obtained better responses in simple prompts compared to the overall graph, this increase has been in Slightly Better responses, and it has lost some quality in the Much Better responses. Unlike ChatGPT, which, although it has lost (**-1.5%**), its Much Better responses have increased by **+2.2%**.

3.5.3 <u>Overall Classification Category</u>, (hyperspecific prompts)



We can see how **ChatGPT** had the greatest success rate with **hyperspecific prompts, at 52.2%**. **+4.4%** better than **simple prompts**.

On the other hand, **Bard** had a **loss** of (**-4.3%**) **compared to simple prompts**, demonstrating that it has struggled with this type of prompt.

What's more, **ChatGPT** has been selected **more times than Bard and ties combined**: **52.2% > 47.8%**

And since the tie percentage remains the same, 21.7%, this shows that **ChatGPT has been quite successful** in this type of prompt.

3.5.5 Conclusion Classification

In conclusion, we can say that **ChatGPT** has **outperformed Bard**, both in general terms and in simple and hyperspecific prompts.

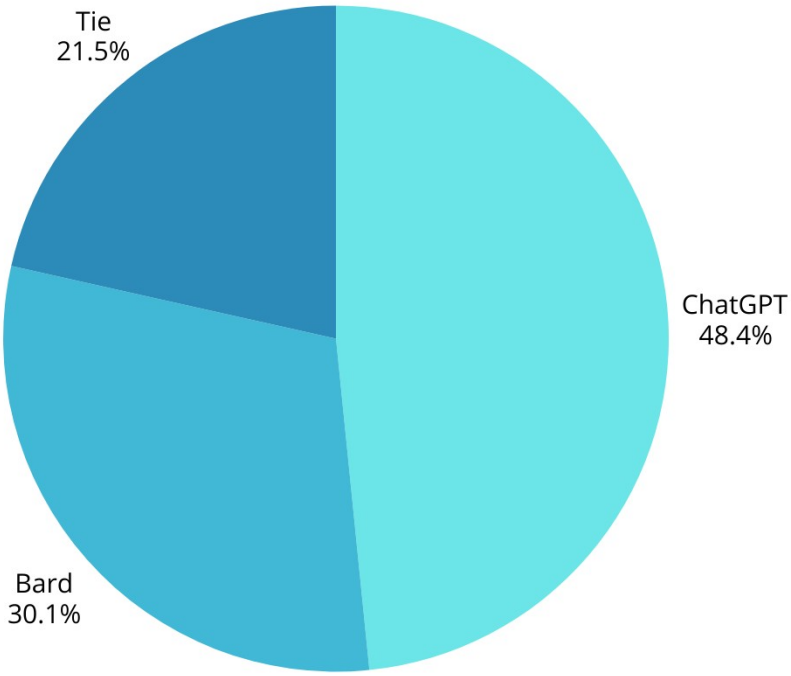
While overall, **ChatGPT** had a response rate of **49.3%**, selected 20.3% more times than Bard.

In **simple prompt ChatGPT dropped** by (**-1.5%**).

However, although it saw this overall decrease in responses **in simple prompts**, **its Much Better Responses increased by +2.2%**, while **Bard's Much Better Responses decreased by** (**-1.5%**). This means that despite **Bard** being selected overall in simple prompts by **+1.4%** compared to its original value of **29%**, its **quality of Much Better responses decreased**, while ChatGPT's increased.

As for **hyperspecific prompts**, **ChatGPT** has been selected **+4.4%** over simple prompts, while **Bard** has decreased (**-4.3%**) over simple prompts, suggesting that **Bard has not had great success in responding to this prompt**.

3.6 Overall and Detailed Closed QA Category

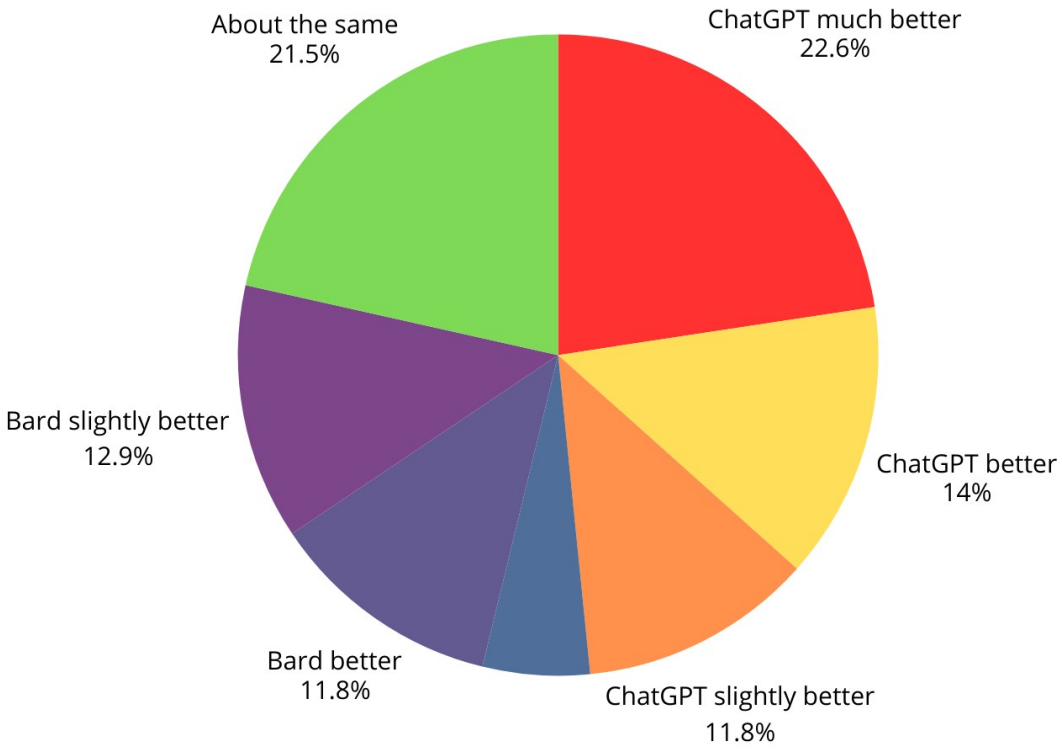


As we can see, **ChatGPT** provides the best responses for the **Closed QA Category** (global), with **48.4%**.

On the other hand, we can see that **Bard Model was selected 30.1%**.

And **21.5%** of the time, **both models have been similar,** (about the same).

This means that **ChatGPT** was selected **18.3%** more times than **Bard.**



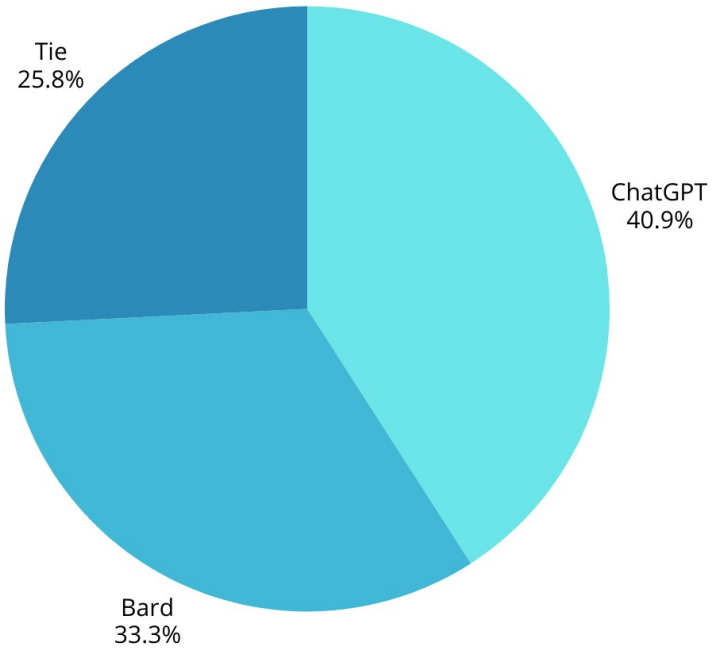
And if we look in detail, we can see how **ChatGPT is superior to Bard in practically all responses**.

ChatGPT much better responses (**22.6%**) vs. **Bard much better responses** (**5.4%**): **17.2% times more.**

ChatGPT better responses (**14%**) vs. **Bard much better responses** (**11.8%**): **2.2% times more.**

Only in **slightly better responses was Bard selected 1.1% more**, that is, barely noticeable.

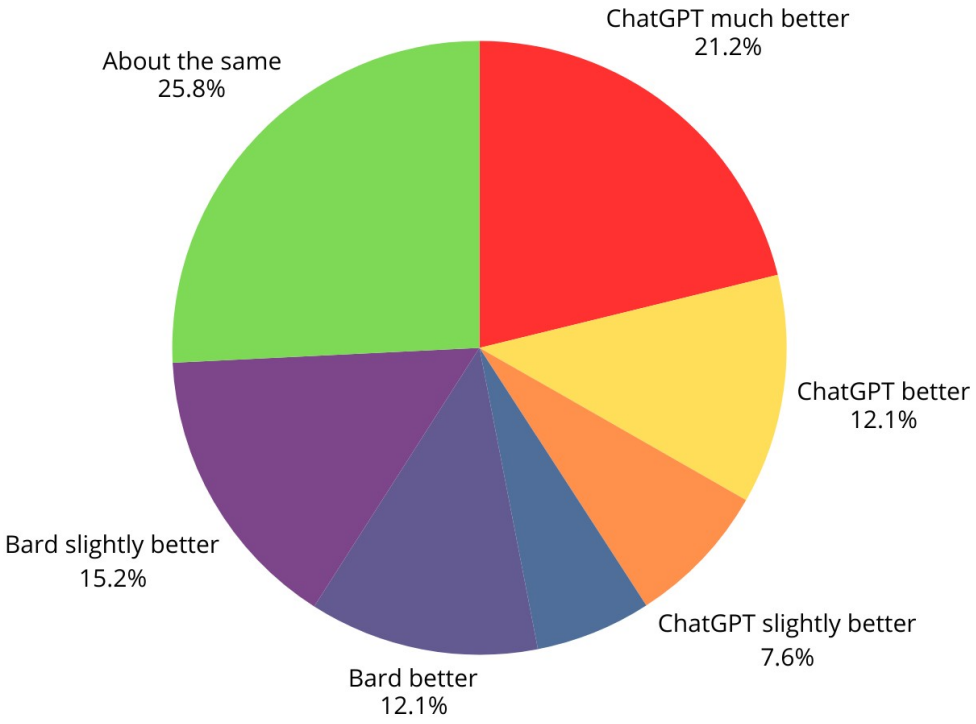
3.6.2 Overall and Detailed Closed QA Category, (simple prompts)



As for **simple prompts** we can see that **ChatGPT** continues to lead, with **40.9%**, **decreasing** by (**-7.5%**) over overall graphic.

Bard, on the other hand, has **increased** by **+3.2%**, while the **ties** have also **increased** by **+4.3%**.

To find out if this is because Bard performed better on simple prompts or because ChatGPT performed worse on simple prompts, we'll look at the graph in detail.

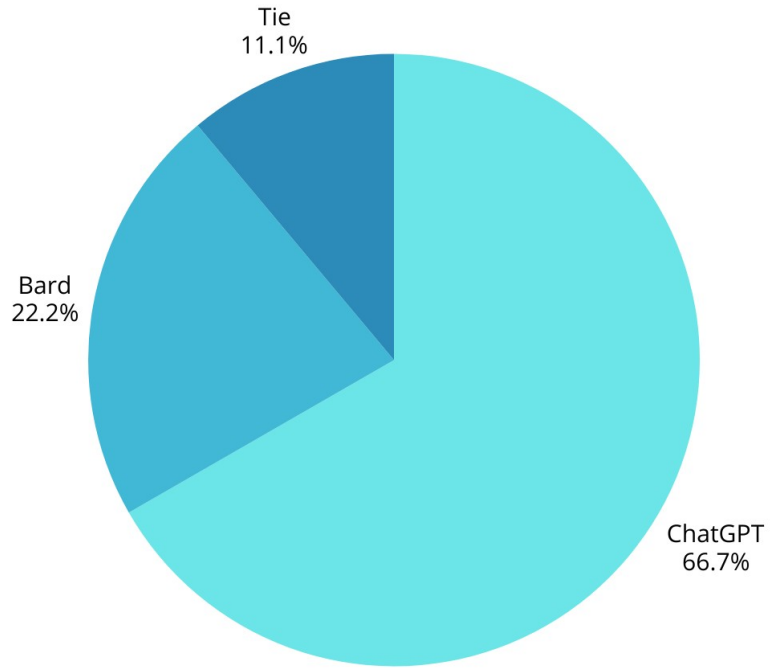


What stands out most in the graph is how **ChatGPT** has **lost the most in slightly better responses**, with a significant loss of (**-4.2%**).

On the other hand, **bard's slightly better responses have increased** by **+2.3%** and **ties** by **+4.3%**.

This means that rather than ChatGPT performing worse, it's more likely that **Bard's model performed better with simple prompts in this category than with other types of prompts**.

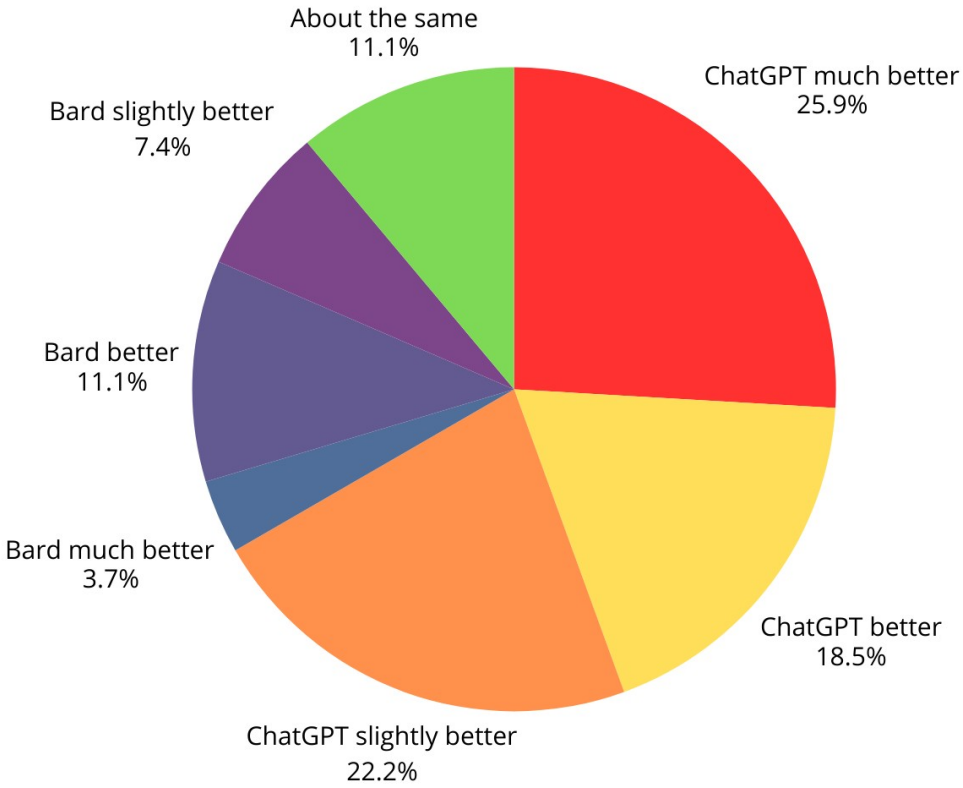
3.6.3 Overall and Detailed Closed QA Category, (hyperspecific prompts)



And our theories are confirmed when we see how, in the case of **hyperspecific prompts**, **ChatGPT** is selected **66.7%** of the time, with a **+25.8%** **increase compared** to **simple prompts**.

At the same time, (compared to simple prompts**), Bard decreased** (**-11.1%**) and **ties** (**-14.7%**).

This implies that ChatGPT has been far superior and that the Bard model has been far inferior, with unsuccessful responses.



Specifically, if we compare the graph, we'll see that all **ChatGPT types of responses have increased** significantly **compared to simple prompts.**

Furthermore, just compare the **ChatGPT Much Better response rate, 25.9%,** to the **Bard Much Better response rate, 3.7%**.

3.6.5 Conclusion Closed QA

In conclusion, we can say that the **ChatGPT model was superior** to the Bard model in the **Closed QA category**.

The **ChatGPT model was chosen 48.4%** of the time, while **Bard was selected 30.1%**.

This means that **ChatGPT was selected 18.3% more** times than Bard.

Looking at the performance of both models solely **on simple prompts**, we can see that **ChatGPT decreased by** (**-7.5%**), while **Bard increased** by **+3.2%**, and **ties** by **+4.3%**.

In this case, **ChatGPT's greatest loss comes from Slightly Better responses**, with a loss of (**-4.2%**), while **Bard's Slightly Better responses increased** by **+2.3%** and **ties** by **+4.3%**.

This suggests that rather than ChatGPT losing effectiveness with this type of prompt, **Bard has actually improved its quality compared to hyperspecific prompts**.

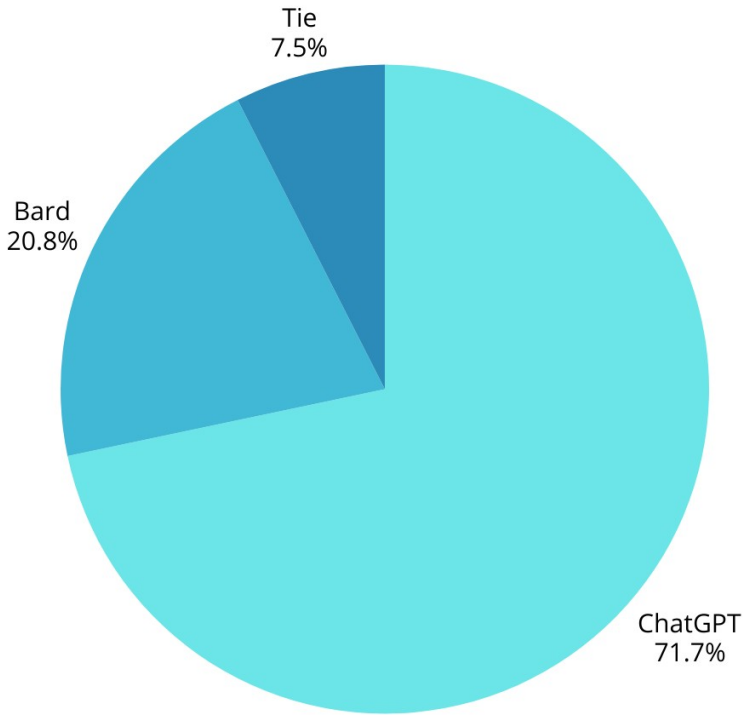
And in the **hyperspecific prompts**, **ChatGPT** is selected **66.7%** of the time, with a **+25.8% increase** compared to simple prompts.

At the same time, **Bard decreased** (**-11.1%**) and **ties** (**-14.7%**), both compared to simple prompts.

Furthermore, just compare the **ChatGPT Much Better response rate**, **25.9%**, to the **Bard Much Better response rate**, **3.7%**.

So specifically for the **hyperspecific prompts Closed QA category, Bard is quite inefficient**.

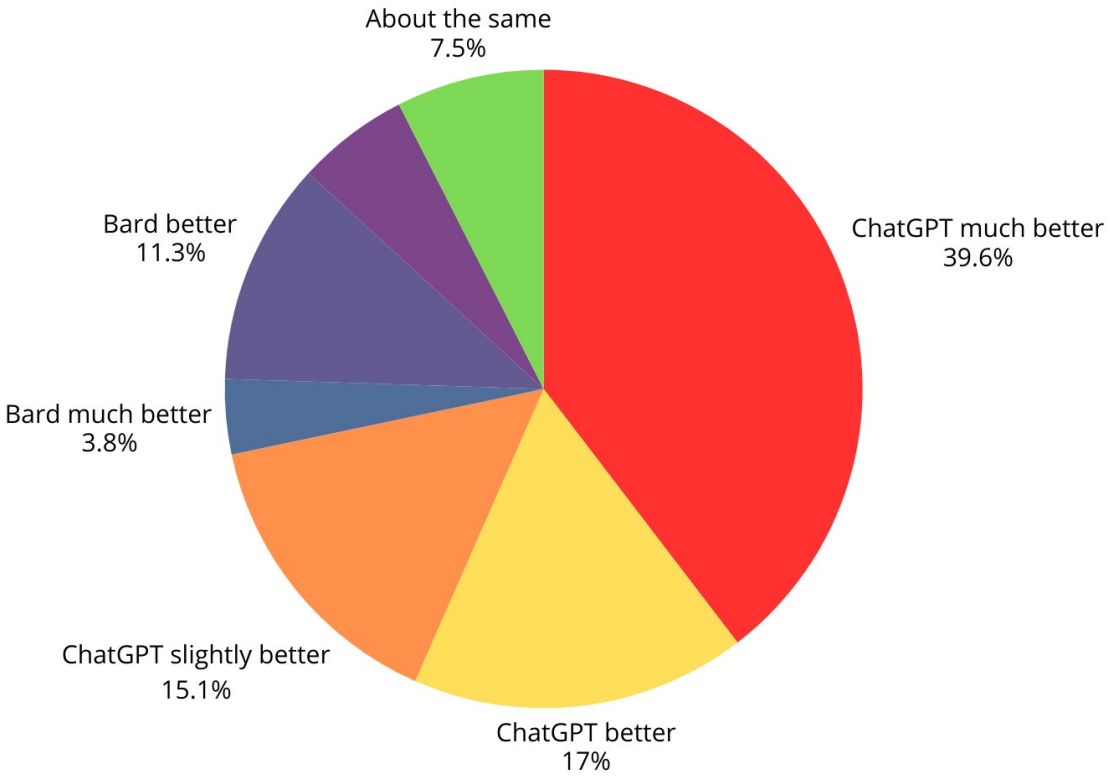
3.7 Overall and Detailed Coding Category



For this category, we can see how **ChatGPT is infinitely superior to Bard**.

ChatGPT leads with 71.7%, while **Bard is selected 20.8%**.

This means that **ChatGPT is selected 50.9% more** often than Bard.

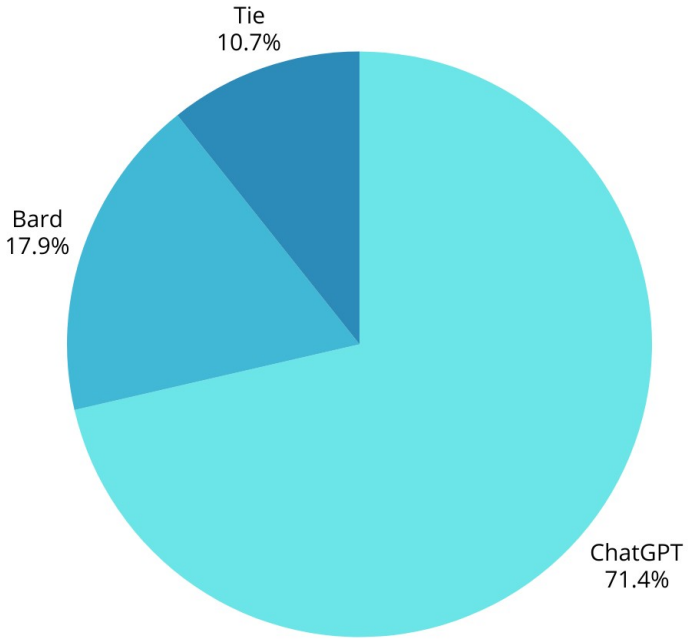


Furthermore, we can see how **ChatGPT has an abysmal 39.6% Much Better response rate**.

Bard Much Better is just 3.8%.

It's clear that ChatGPT is much better at programming than Bard.

3.7.2 Overall Coding Category, (simple prompts)

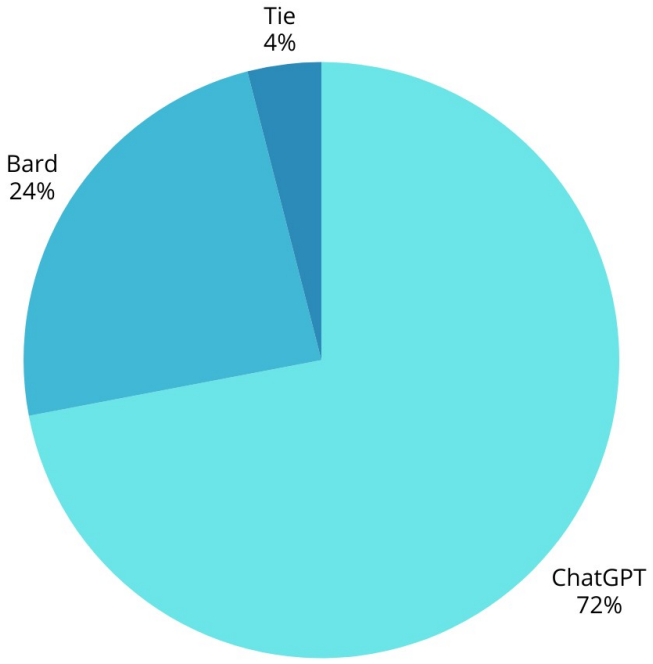


As for **simple prompts** we can see that **ChatGPT** continues to lead, with **71.4%**, **decreasing only** (**-0.3%**).

At the same time, **Bard also decreased** (**-2.9%**) and **ties increased** by **+3.2%**.

These are hardly significant changes, and it can be said without a doubt that **ChatGPT is far superior**.

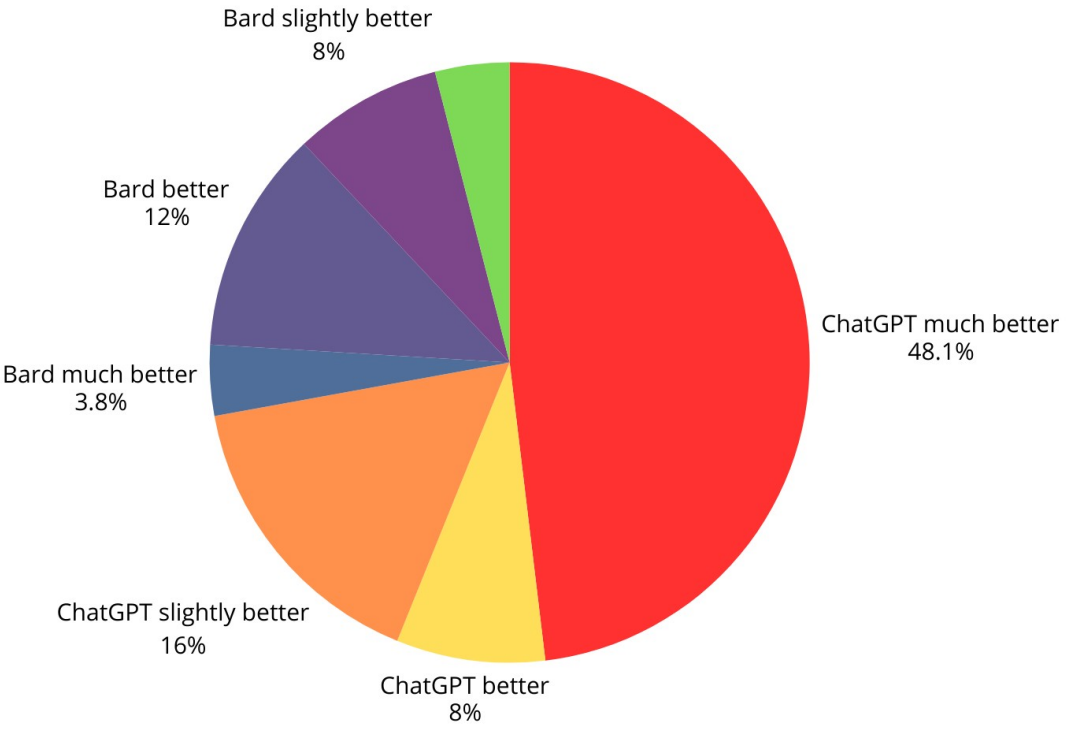
3.7.3 Overall and Detailed Coding Category, (hyperspecific prompts)



For **hyperspecific prompts**, **ChatGPT has increased** by just **+0.6% compared to simple prompts**.

However, **Bard has increased** by **+6.1%**, (over simple prompts**),** which shows that for programming, **Bard is better at hyperspecific prompts than simple prompts**.

Ties have decreased (**-6.7%**).



We can see how **ChatGPT leads with 48.1% Much Better Responses**, compared to **Bard's 3.8% Much Better responses**.

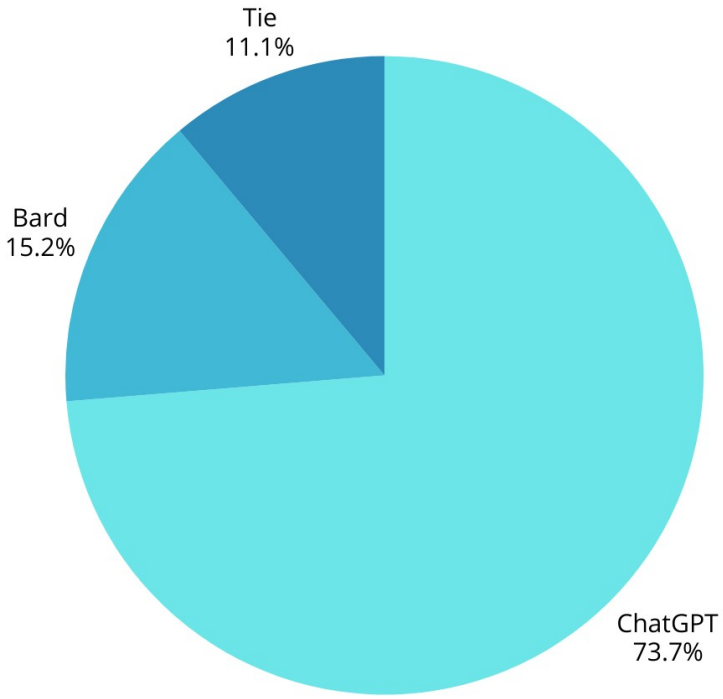
3.7.4 Conclusion Coding

In conclusion, **ChatGPT is far superior to Bard in programming**, with a score of **71.7%** for general programming.

This can be seen as the **models barely match**, with only **7.5% of the models providing the same or similar answers**, meaning the **responses are often very different between the two artificial intelligence models**.

Although **Bard yields unsatisfactory answers in the programming area**, a **significant increase of +6.1% was observed in hyperspecific prompts** compared to simple prompts, **which means it has some problems with simple prompts**.

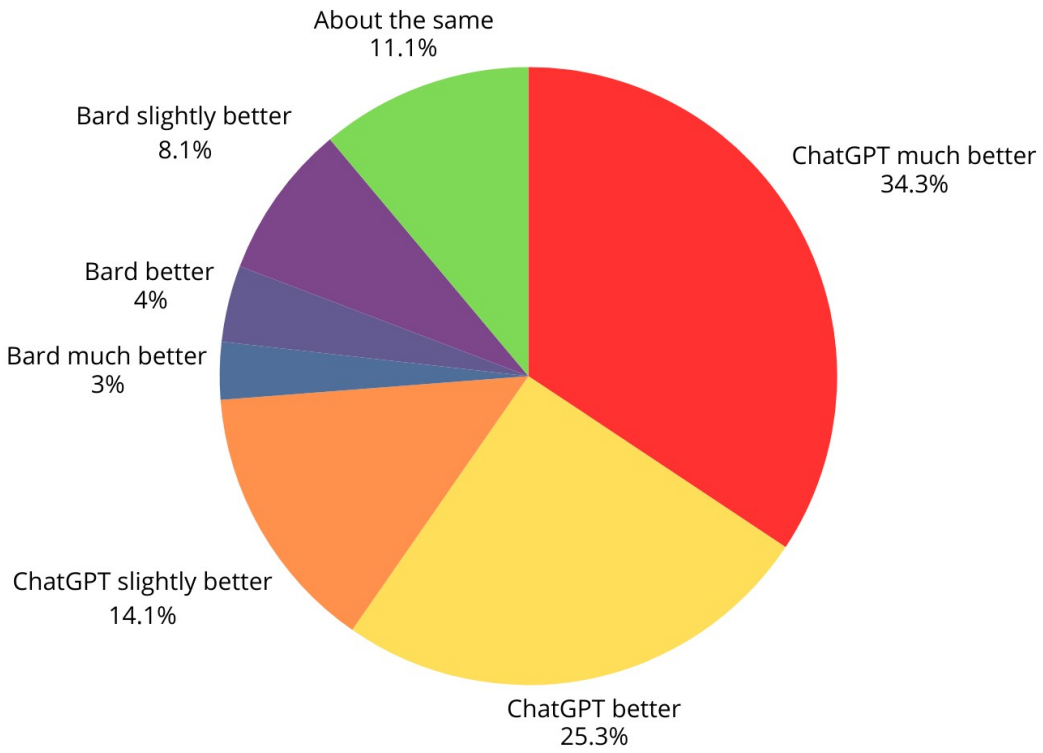
3.8 Overall and Detailed Creative Writting Category



As we can see, **ChatGPT** provides the best responses for the **Creative Writting Category** (global), with **73.7%**.

This is outrageous given that if we add up Bard's responses and the ties, we get **26.3%**, which is less than half of ChatGPT

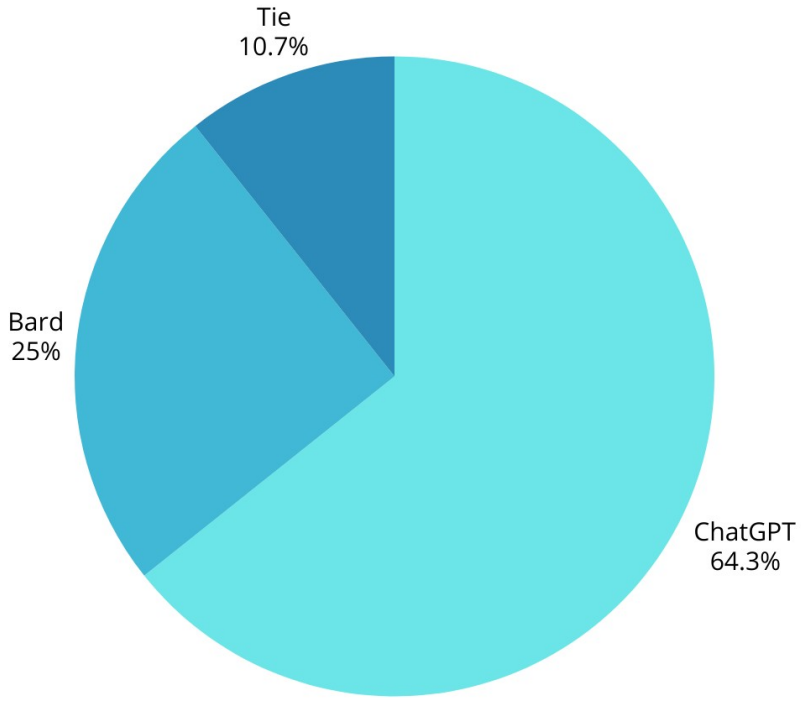
In any case, one thing is clear: **ChatGPT was selected 47.4% more than Bard**.



We can also see how **ChatGPT's Much Better response** rate accounts for **34.3%** of the responses, while **Bard's Much Better response** rate accounts for only **3%**.

This means that **ChatGPT provides very satisfactory responses for this category** compared to Bard.

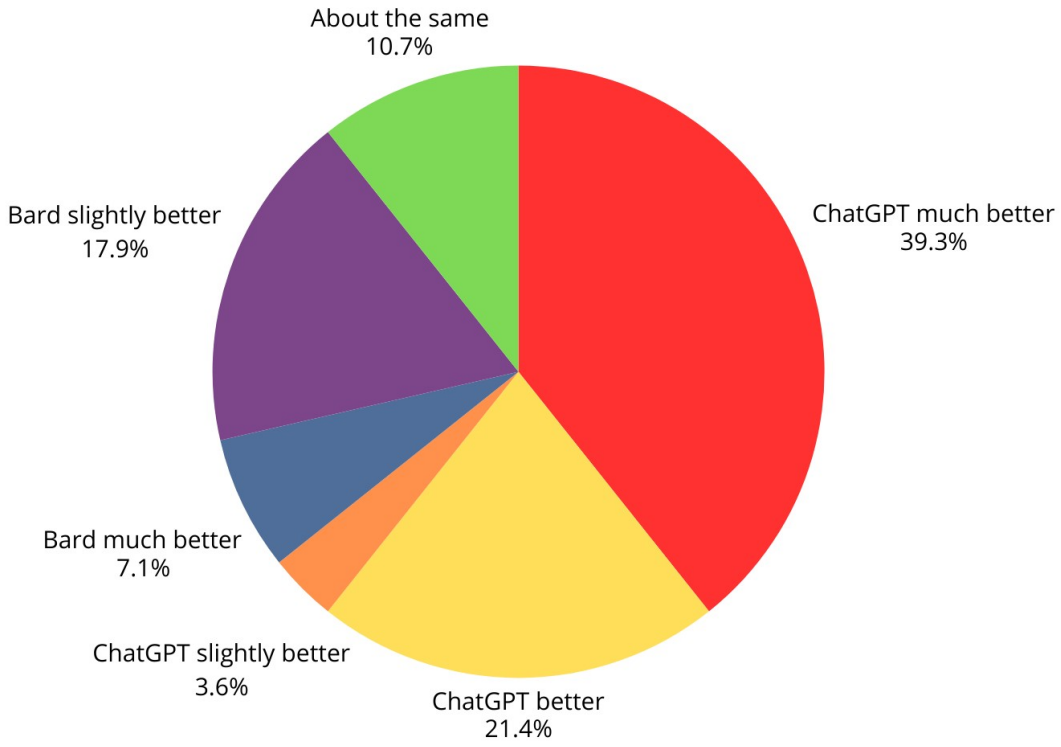
3.8.2 Overall and Detailed Creative Writting Category, (simple prompts)



As for **simple prompts** we can see that **ChatGPT** continues to lead, with **64.3%**.

At the same time, **Bard increased in simple prompts** by **+9.8%**.

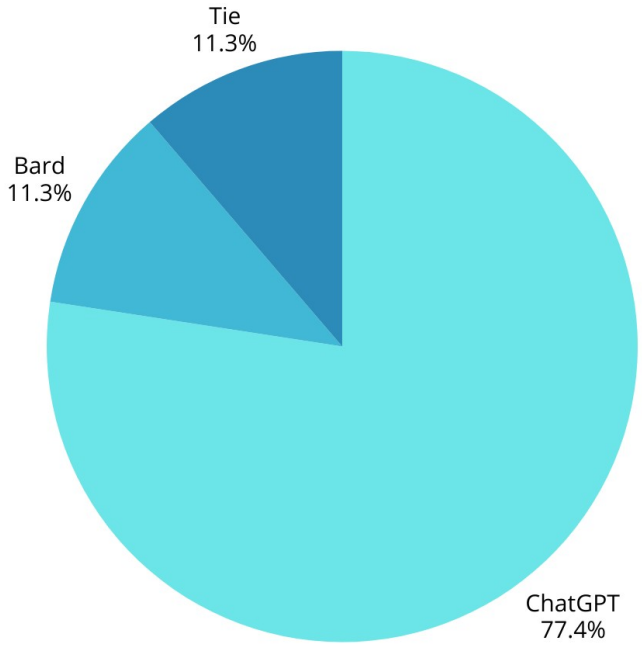
This means that while ChatGPT remains comfortably in the lead, it's possible to draw conclusions from the previous graph that Bard has struggled to provide successful responses to other types of prompts.



As for **simple prompts** we can see that **ChatGPT Much Better** responses continues to lead with **39.3%,** which is an **increase of +5%**.

However, a curious fact is that **Bard was only Much Better, 7.1%** or **Slightly Better, 3.6%**. It wasn't just **Better** than ChatGPT at no time.

3.8.3 Overall Creative Writting Category, (hyperspecific prompts)



For **hyperspecific prompts**, **ChatGPT has increased +13.1% over simple prompts**.

However, **Bard has decreased** (**- 6.1%**), which shows that for Creative Writting, **Bard has a hard time at hyperspecific prompts**.

3.8.4 Conclusion Creative Writting

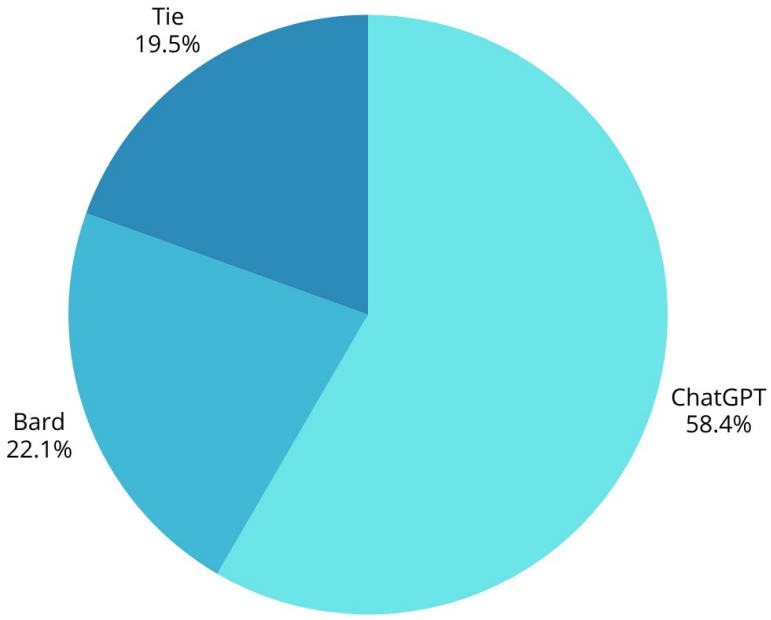
In conclusion, **ChatGPT is far superior to Bard in Creative Writting**, with a score of **73.7%**.

It can also be added that **Bard struggled** to provide successful responses **to hyperspecific prompts**, with a **drop of** (**-6.1%**) **compared to simple prompts**, where it had a **higher success rate**, at **25%**.

This suggests that while ChatGPT is better in all aspects for this category and **Bard struggles to respond efficiently to both types of prompts**, **hyperspecific prompts are even more difficult for it**.

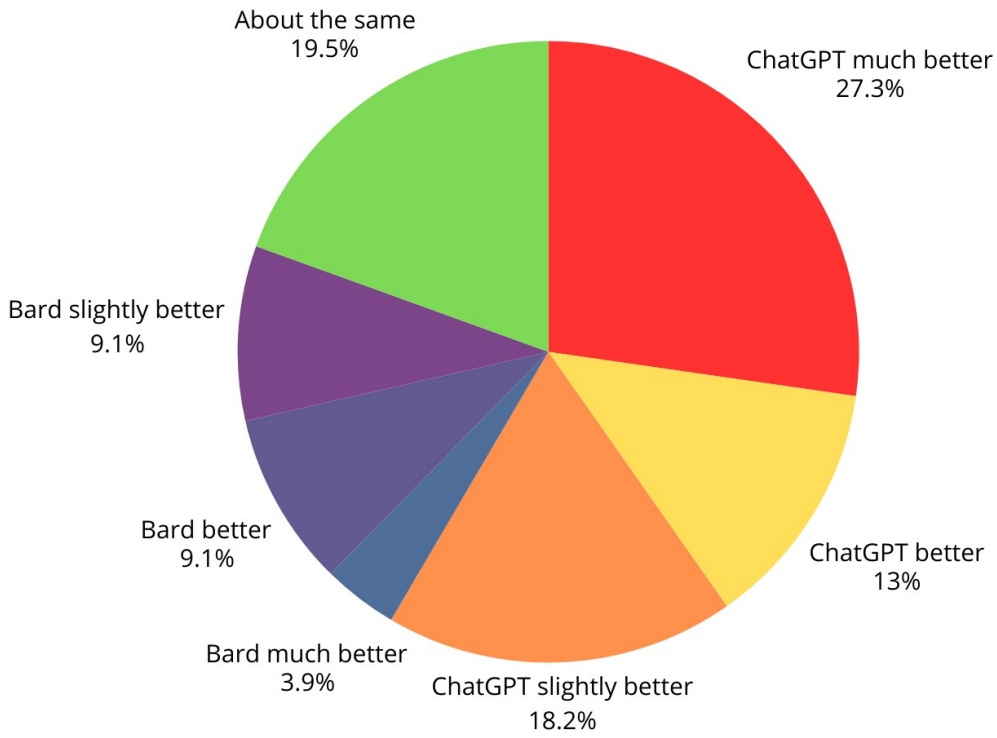
As for **ties**, the percentage was quite low, at **11.1%** for overall, suggesting that **responses tend to be quite different in quality between models**.

3.9 Overall and Detailed Extraction Category



For this category we can see how **ChatGPT** gets **58.4%** of better answers

Bard gets 22.1% and **ties** have the percentage of **19.5%**

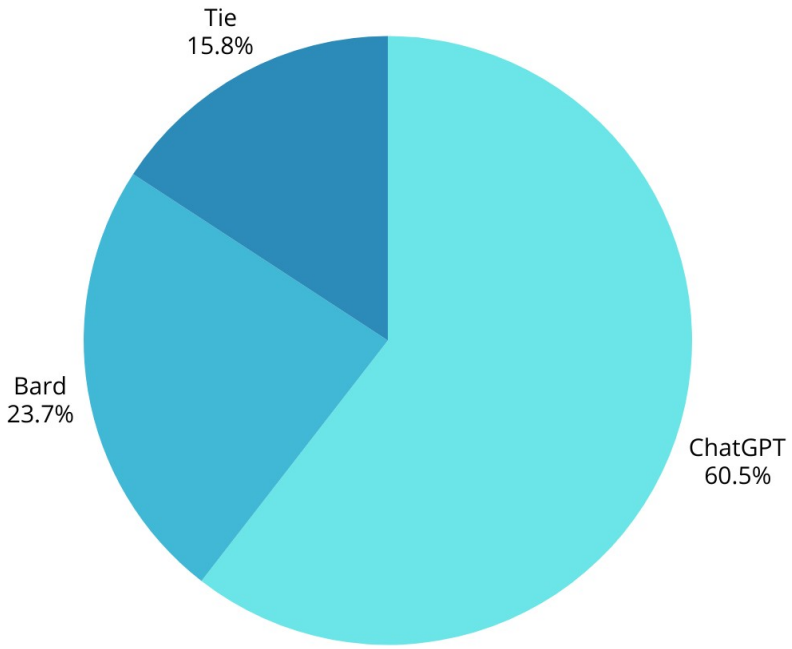


As we can see, **ChatGPT's Much Better** responses lead with **27.3%**, which is **already higher than the sum of Bard's Slightly, Better, and Much Better responses**.

27.3% > 22.1%

This suggests that **ChatGPT provides very satisfactory responses**.

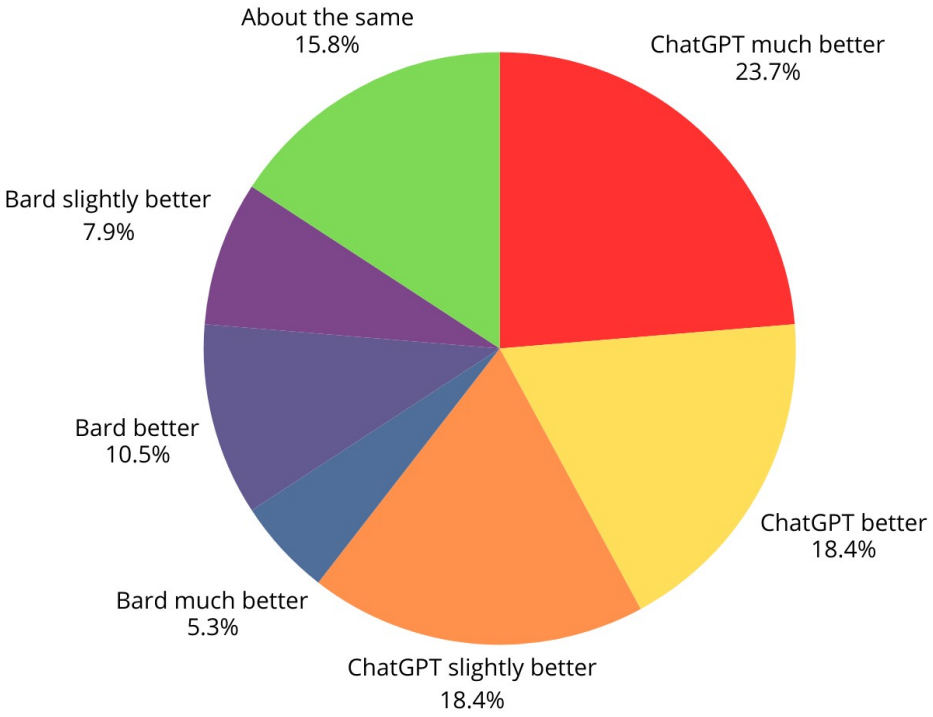
3.9.2 Overall and Detailed Extraction Category, (simple prompts)



As for **simple prompts** we can see that both ChatGPT and Bard increase relative to the overall prompt graph.

ChatGPT increased by **+2.1%** while **Bard increased** by **+1.6%.**

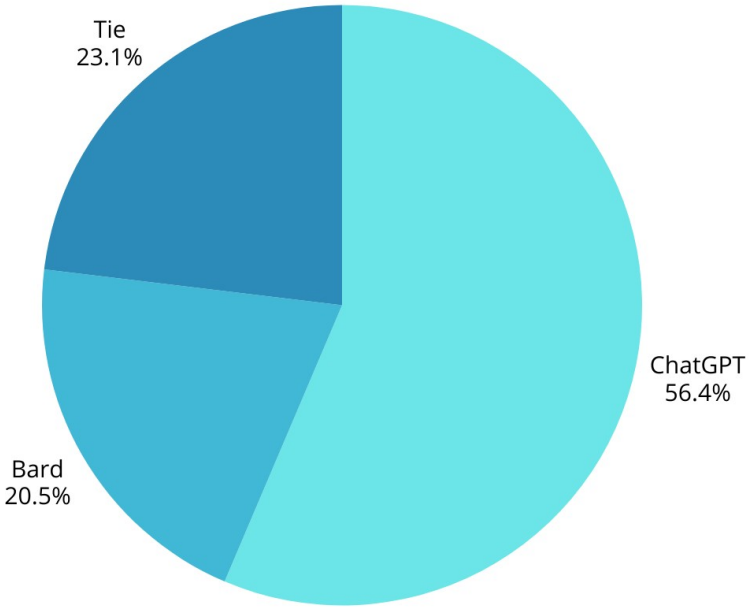
Ties decreased (-3.7%) compared to the overall graph**, suggesting that responses were less similar.**



Looking at the details, we can see that **ChatGPT's Much Better** responses have **dropped** by (**-7.3%**), which is significant.

Meanwhile, **Bard's Much Better** responses have **increased** by **+1.4%**.

3.9.3 Overall and Detailed Extraction Category, (hyperspecific prompts)

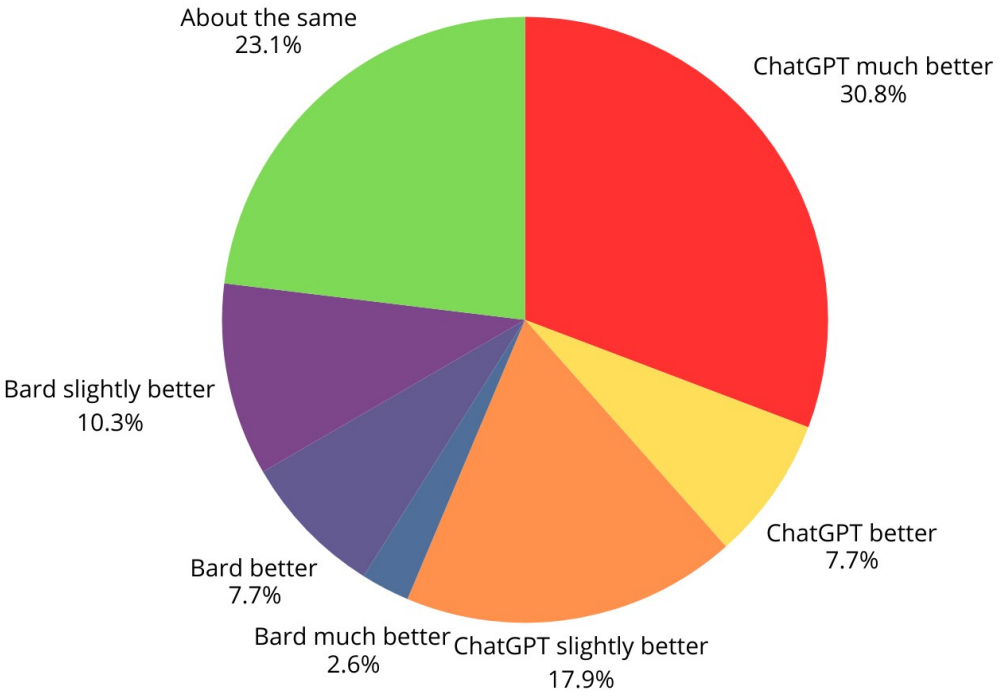


For **hyperspecific prompts**, **ChatGPT ha decreased (-4.1%) compared to simple prompts**.

Bard has also **decreased** by (**-3.2%**).

However, **draws** have **increased** by **+7.3%**.

Exactly what **was lost by both models,** was **added to the ties**, suggesting that compared to simple prompts, **both models had more similar responses**.



However, **ChatGPT**, despite having lost (**-4.1%**) **compared** to **simple prompts**, as we can see in the graph, has seen a **+7.1% increase in Much Better Responses**, indicating that it has actually **succeeded with hyperspecific prompts more than with simple prompts.**

This contrasts with **Bard's Much Better Responses**, which has **decreased** (**-2.7%**) compared to simple prompts.

This suggests that **Bard has had difficulties with this type of prompt.**

3.9.4 Conclusion Extraction

In conclusion, **ChatGPT is far superior to Bard in Extraction**, with a score of **58.4%, while Bard had 23.7%.**

It can also be added that **Bard struggled** to provide successful responses **to hyperspecific prompts**, with a **drop of** (**-3.2%**) **compared to simple prompts**, of which, (**-2.7%**) was loss of **Much Better Responses**.

And **ChatGPT**, even though it had a **loss** of (**-4.1%**) **compared to simple prompts**, had a gain of **+7.1% in Much Better Responses**.

As for **simple prompts, ChatGPT increased by +2.1%** while **Bard increased by +1.6%**.

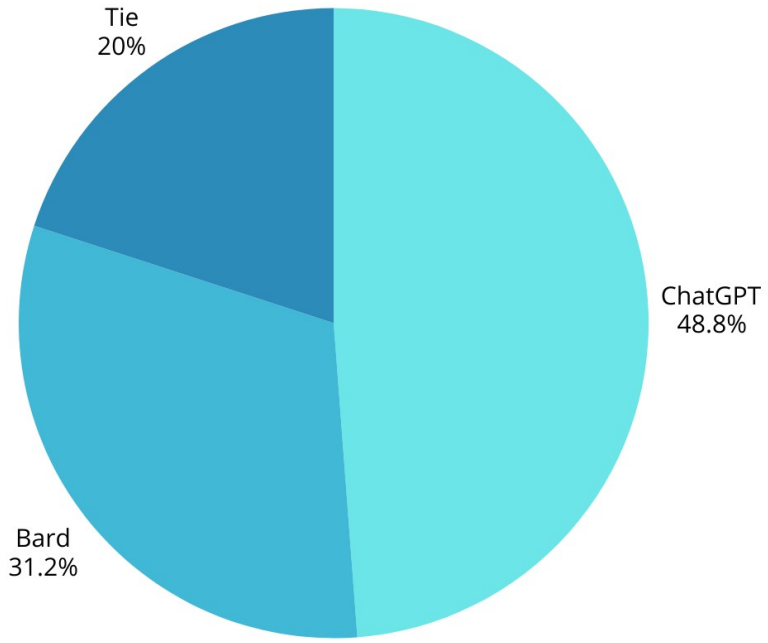
ChatGPT's Much Better responses have **dropped** by (**-7.3%**), **on simple prompts** which is significant.

Meanwhile, **Bard's Much Better responses have increased** by **+1.4%**.

Meanwhile, in **simple prompts**, **ties decreased** (**-3.7%**) compared to the overall graph, suggesting that responses were less similar.

This makes it clear that **Bard had fewer problems with simple prompts than with hyperspecific prompts.**

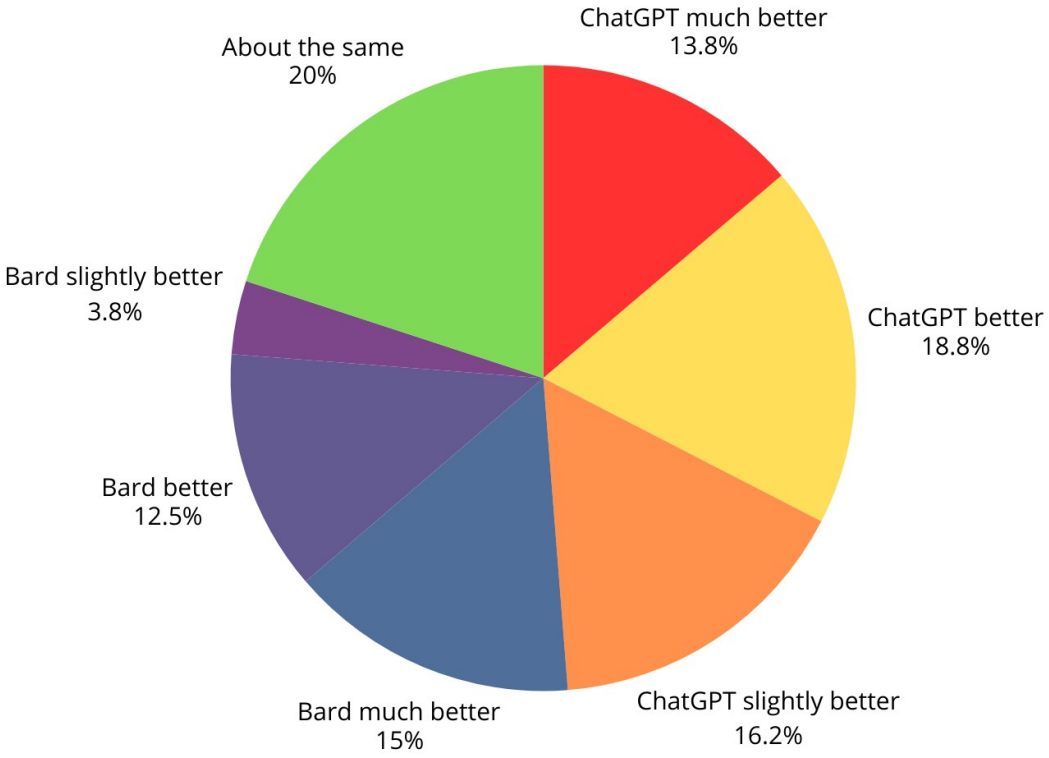
3.10 Overall and Detailed Mathematical Reasoning Category



For this category we can see how **ChatGPT** gets **48.8%** of better answers

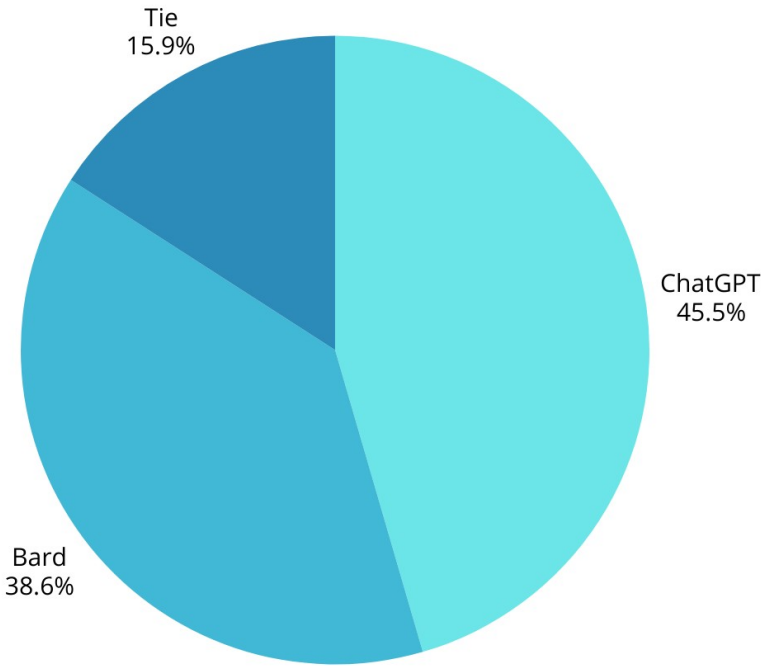
Bard gets 31.2% and **ties** have the percentage of **20%**

This tells us that **ChatGPT was selected 17.6%** higher.

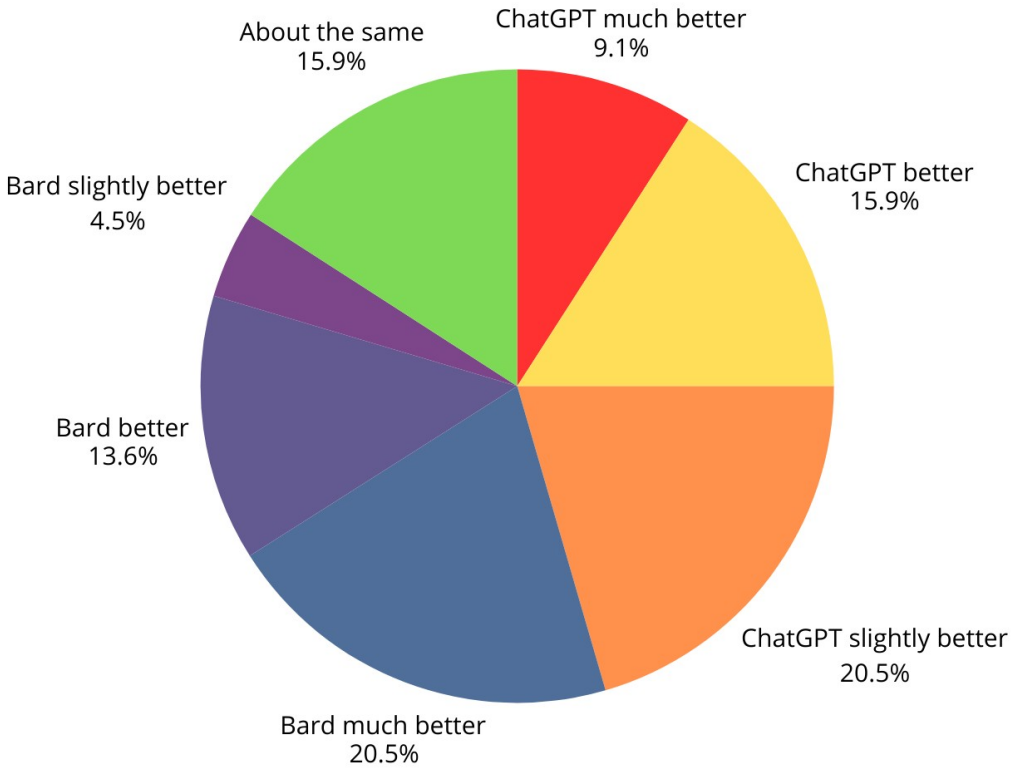


However, despite this, we can see that **Bard** had more **Much Better responses**, with **15%**, than **ChatGPT**, with **13.8%**.

3.10.2 Overall and Detailed Mathematical Reasoning Category, (simple prompts)



For the simple prompt we can see that **ChatGPT decreased** (**-3.3%**) and **Bard increased** **+7.4%** as well as **ties** decreased (**-4.1%**).

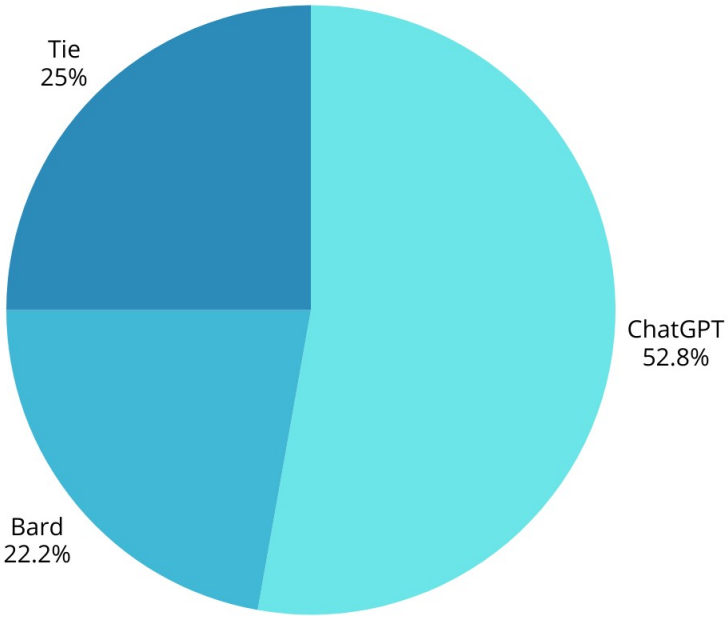


Furthermore, we can see how **ChatGPT struggled** to create **excellent responses with simple prompts**, i.e., **Much Better Responses**, with **9.1%**.
On its counterpart, **Bard's Much Better Responses reached** a surprising **20.5%**.

Also, we can see that the only clear difference where **ChatGPT** is superior is **in Slightly Better responses, with 20.5% compared to Bard's 4.5%**.

This suggests that **although ChatGPT obtained more winning responses in simple prompts** (by just **6.9%**), **Bard's responses were of higher quality, as can be seen in the 20.5% Much Better Responses** percentage.

3.10.3 Overall and Detailed Mathematical Reasoning, (hyperspecific prompts)

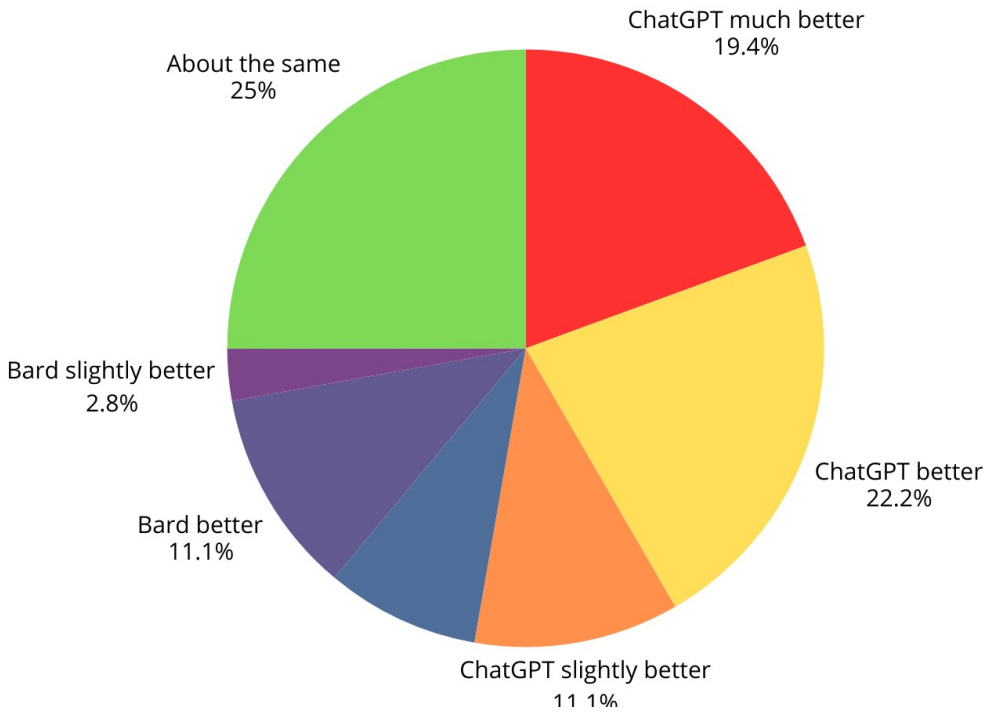


For **hyperspecific prompts**, **ChatGPT ha increased +7.3% compared to simple prompts**.

Bard has also **decreased** (**-16.4%**).

However, **draws** have **increased** by **+9.1%**.

This means that **ChatGPT** has **performed better in terms of hyperspecific prompts** for this **category**, and **Bard** has had quite a bit of **difficulty**.



This can be seen in the detailed graph, where we see that **ChatGPT Much Better** Responses has increased by **+10.3% compared to simple prompts**, while **Bard Much Better** Responses has **decreased** by (**-12.2%**), resulting in a total of **8.3% Much Better** Responses.

This confirms the fact that **Bard has struggled and declined in quality for this type of prompt in the Mathematical Reasoning category.**

3.10.4 Conclusion Mathematical Reasoning

In conclusion, **ChatGPT is overall superior in Mathematical Reasoning**, with a score of **48.8%, while Bard had 31.2%.**

For the simple prompt, **ChatGPT decreased** (**-3.3%**) and **Bard increased** **+7.4%** as well.

Furthermore, **ChatGPT struggled** to create **excellent responses with simple prompts**, i.e., **Much Better Responses**, with **9.1%**. On its counterpart, **Bard's Much Better Responses reached** a surprising **20.5%**.

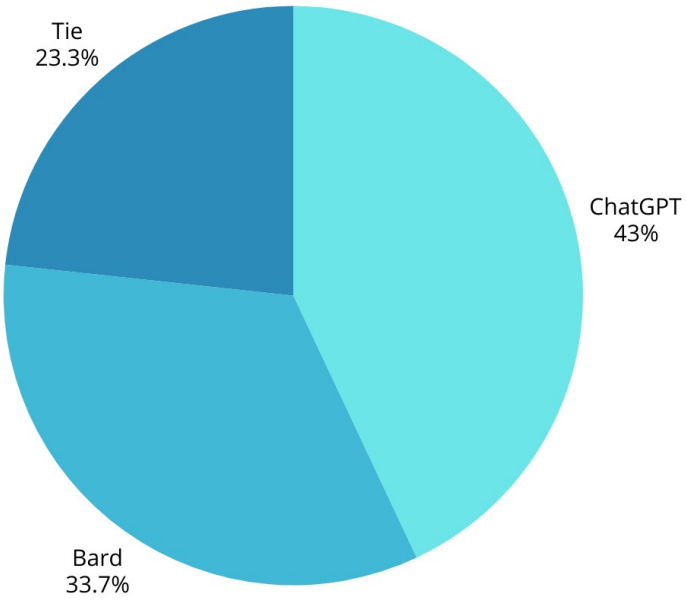
This suggests that **although ChatGPT obtained more winning responses in simple prompts**, **Bard's responses were of higher quality**.

Howeverm for **hyperspecific prompts**, **ChatGPT ha increased** **+7.3%** **overall, compared to simple prompts,** while **Bard decreased** (**-16.4%**).

ChatGPT Much Better Responses increased by **+10.3%.**
On other hand, **Bard Much Better** Responses has **decreased** by (**-12.2%**), resulting in a total of **8.3% Much Better** Responses.

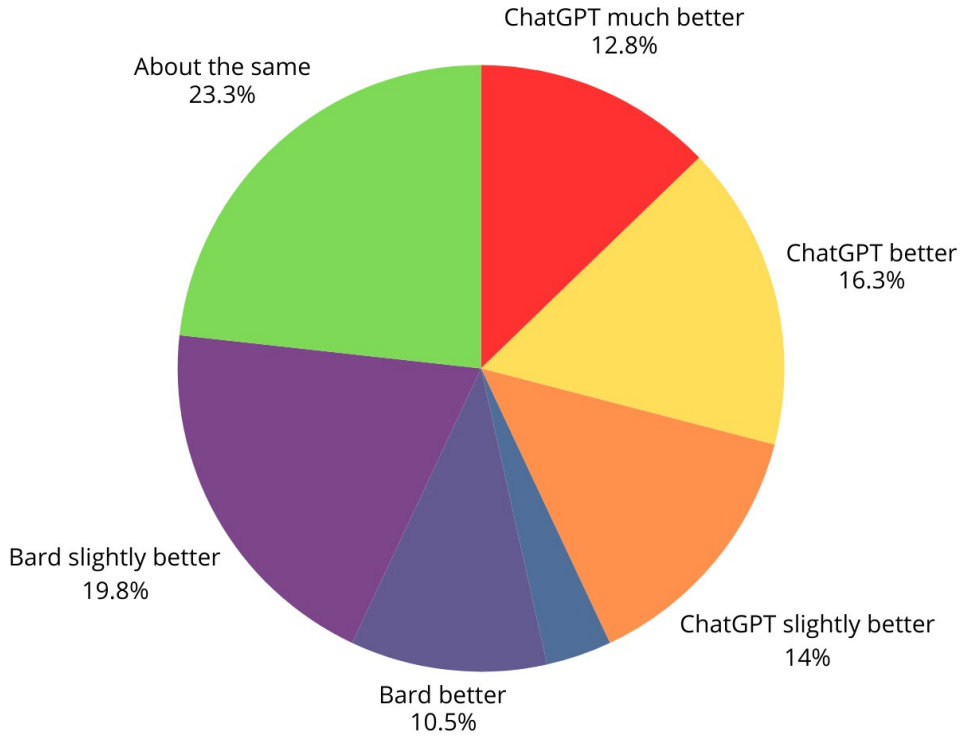
This confirms the fact that **Bard has struggled and decreased in quality for hyperspecific prompts in the Mathematical Reasoning category.**

3.11 Overall and Detailed Open QA Category



For this category, we can see that **ChatGPT's** answers were chosen by users **43%** of the time, while **Brad's answers** were chosen **33.7%** of the time.

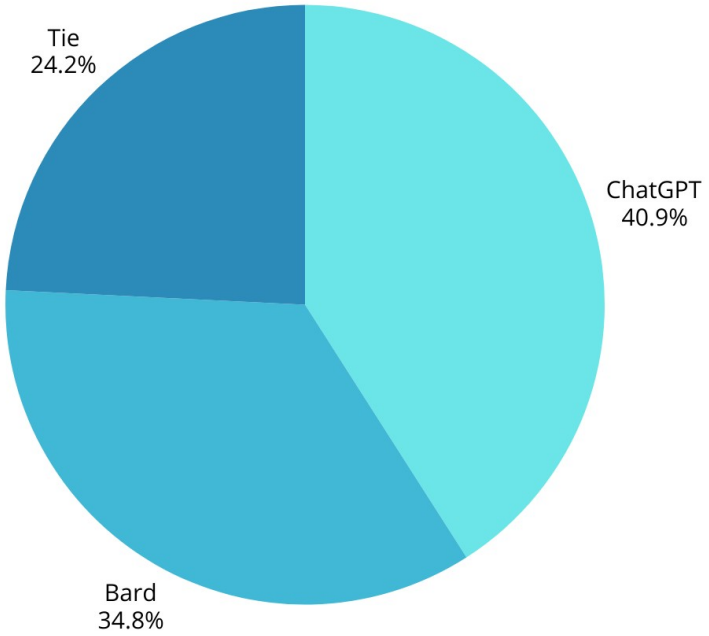
This means that **ChatGPT** was chosen **9.3% more often**.



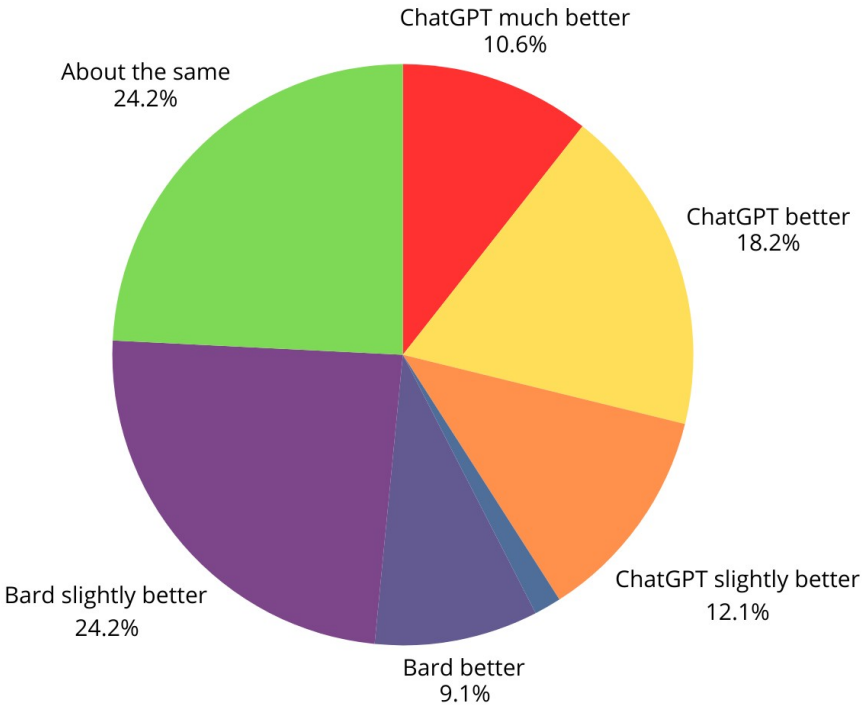
However, while the **9.3%** difference may seem small, if we look closely at the enlarged graph, we see that the **percentages for Slightly and Better responses are similar**, with the **exception of Much Better**, which **ChatGPT** has at **12.8%** while **Bard** has at **3.5%**.

This suggests that the **9.3% difference is largely related to the number of excellent responses ChatGPT received**.

3.11.2 Overall and Detailed Open QA Category, (simple prompts)



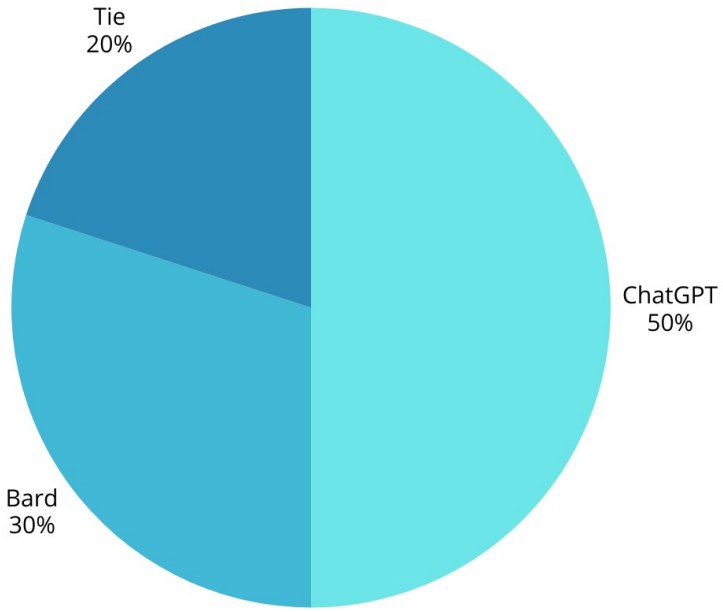
For the simple prompt we can see that **ChatGPT decreased** (**-2.1%**) and **Bard increased** **+1.1%**.



As we can see, there are no significant changes except that **ChatGPT Much Better** has decreased (**-2.2%**), while **Bard Much Better** has decreased also (**-2%**).

However, **Bard Slightly Better** responses have **increased** by **+4.4%**, and **ChatGPT Better** responses have **increased** by **+1.9%**.

3.11.3 Overall and Detailed Open QA, (hyperspecific prompts)

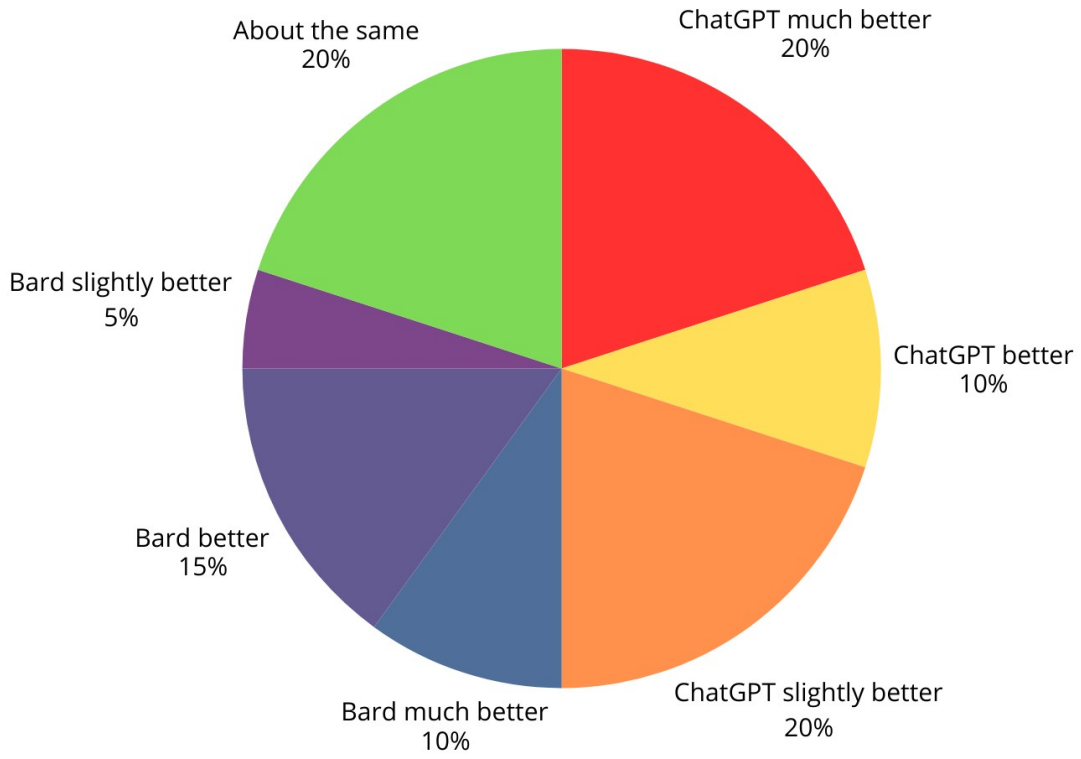


For **hyperspecific prompts**, **ChatGPT has increased +9.1% compared to simple prompts**.

Bard has **decreased** (**-4.8%**).

Draws have also **decreased** (**-4.2%**).

This means that **ChatGPT** has **performed better in terms of hyperspecific prompts** for this **category**, and **Bard** has had quite a bit of **difficulty…**, or not.



As we can see, **ChatGPT Much Better** has **increased** by **+9.4%** **compared to simple prompts**.

However, we note how **Bard Much Better** has **increased** by **+8.5%**, which is a considerable increase **compared to simple prompts**.

In fact, if we **sum Bard Much Better and Better responses from hyperspecific prompts**, we get a **25% increase**, while for **simple prompts, the sum** of these was **10.6%**.

This means that **Bard has obtained more responses with better quality for hyperspecific prompts**, **despite having been selected fewer times compared to simple prompts**, given that Bard Slightly Better simple prompts are 19.2% higher.

3.11.4 Conclusion Open QA

In conclusion, **ChatGPT is overall superior in Open QA**, with a score of **43%, while Bard had 33.7%.**

For the simple prompt, **ChatGPT decreased** (**-2.1%**) and **Bard increased** **+1.1%**.

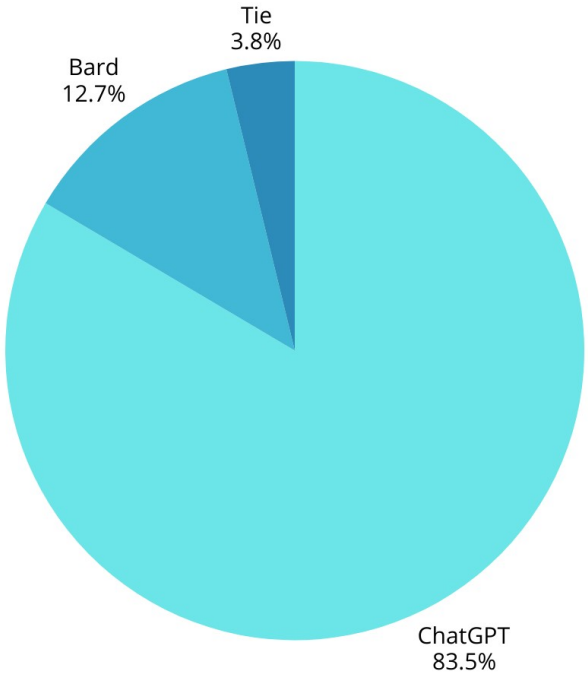
For **hyperspecific prompts, ChatGPT** has **increased** **+9.1%** **compared to simple prompts.**
Bard has **decreased (-4.8%) compared to simple prompts.**
Draws have also decreased (-4.2%) compared to simple prompts.

However, we note how **Bard Much Better** has **increased** +8.5%, which is a considerable increase **compared to Bard Much Better simple prompts**.

In fact, if we **sum Bard Much Better and Better responses from hyperspecific prompts**, we get a **25% increase**, while for **simple prompts**, the **sum** of these was **10.6%**.

This means that **Bard has obtained more responses with better quality for hyperspecific prompts**, despite having been selected fewer times compared to simple prompts, given that Bard Slightly Better simple prompts are **19.2%** higher.

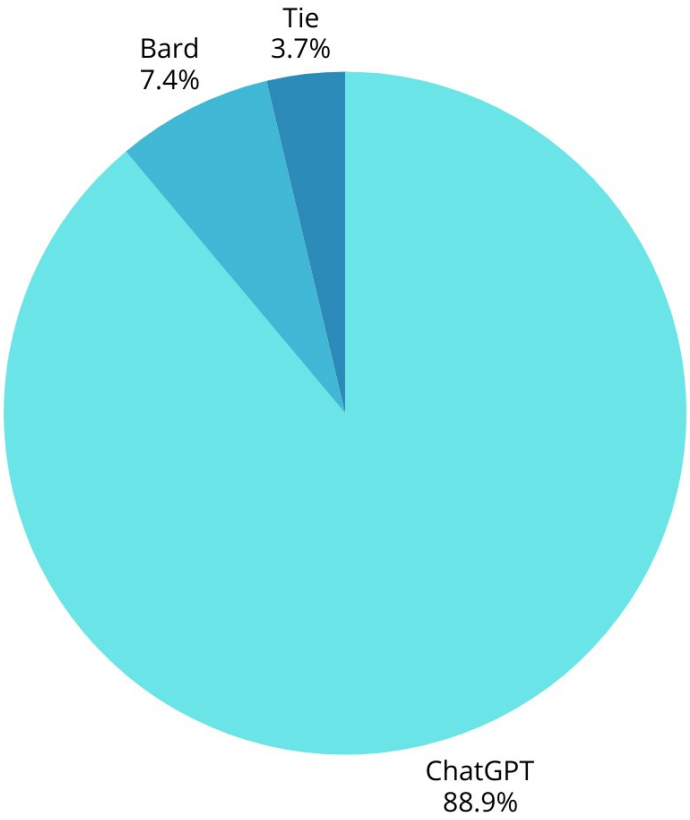
3.12 Overall Poetry Category



For this category, we can see that **ChatGPT's** answers were chosen by users **83.5%** of the time, while **Brad's answers** were chosen **12.7%** of the time.

This represents an absolute majority, making it clear that **Bard is not a good tool in this category**.

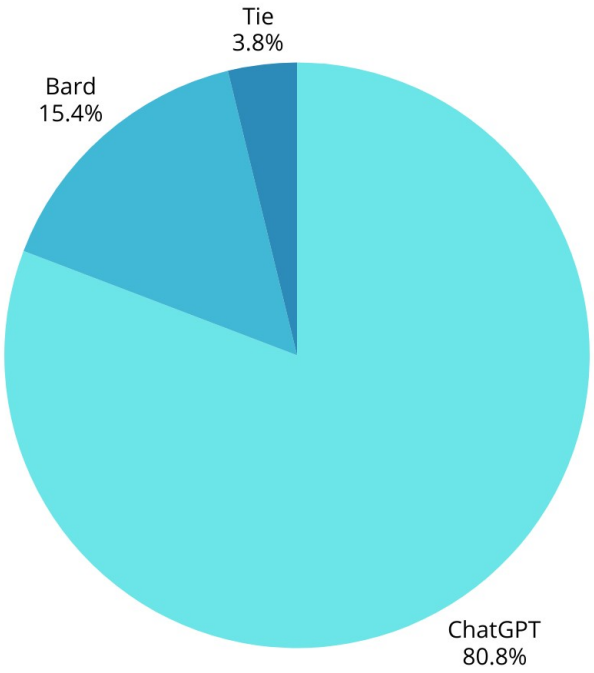
3.12.2 Overall and Detailed Poetry Category, (simple prompts)



For **simple prompts**, we can see how **Brad's performance has decreased** considerably compared to the overall prompts graph, indicating that he has had some issues with this type of prompt, with a **drop of** (**-5.3%**).

On the other hand, **ChatGPT** has **increased** by **+5.4%**, making it clear that simple prompts have not been a problem for him.

3.12.3 Overall Poetry, (hyperspecific prompts)



For **hyperspecific prompts**, **ChatGPT has decreased** (**-8.1%**) **compared to simple prompts**.

Bard has **increased** **+8%** **compared to simple prompts**.

This means that for **hyperspecific prompts**, **Bard has been slightly better than with simple prompts**, even though ChatGPT is the clear winner.

3.12.4 Conclusion Poetry

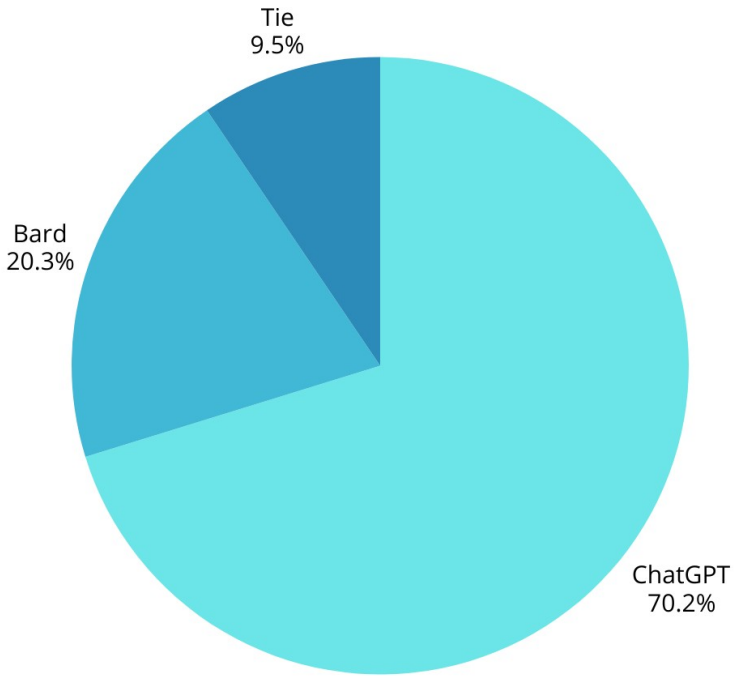
In conclusion, **ChatGPT is far superior in Poetry**, with a score of **83.5%, while Bard had 12.7%.**

For the **simple prompt**, **Bard decreased** (**-5.3%**) and **ChatGPT increased** **+5.4%**.

For **hyperspecific prompts, ChatGPT** has **decreased (-8.1%) compared to simple prompts.**
Bard has **increased** **+8%.**

This means that for **hyperspecific prompts**, **Bard has been slightly better than with simple prompts**, even though ChatGPT is the clear winner.

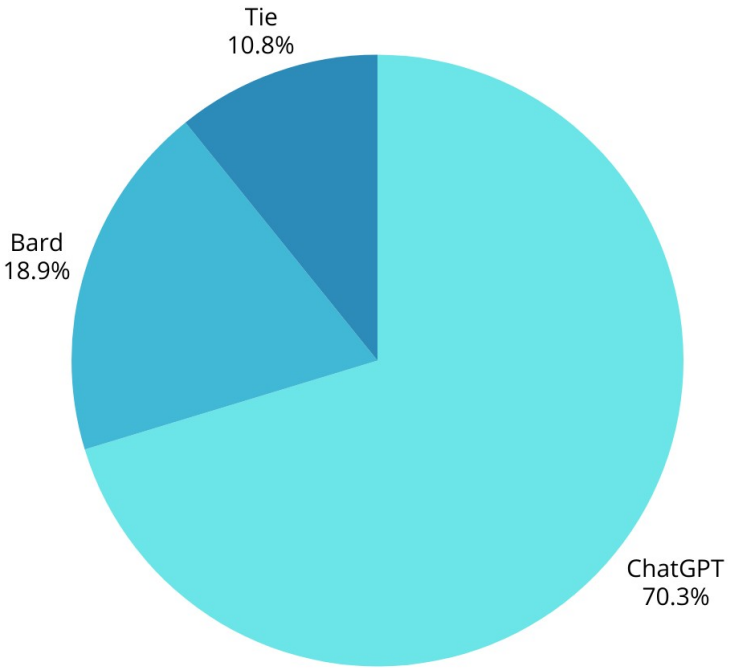
3.13 Overall Rewriting Category



For this category, we can see that **ChatGPT's** answers were chosen by users **70.2%** of the time, while **Brad's answers** were chosen **20.3%** of the time.

This represents an absolute majority, making it clear that **Bard is not a good tool in this category**.

3.13.2 Overall Rewriting Category, (simple prompts)



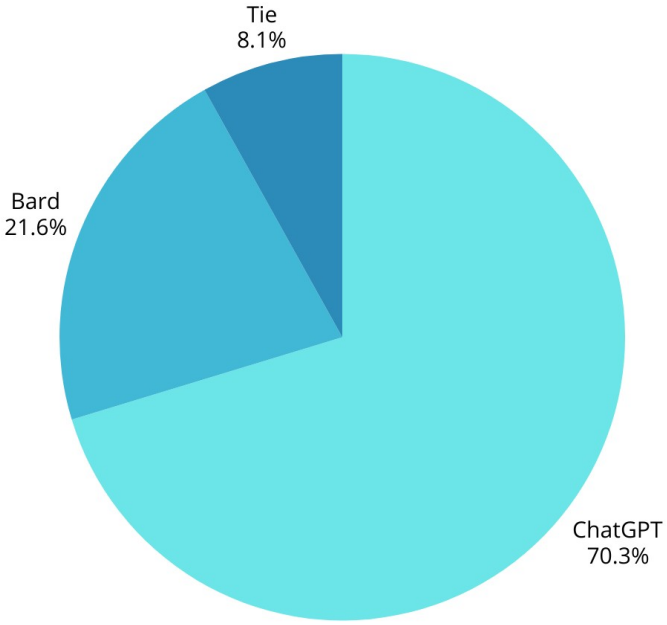
For **simple prompts**, we can see how **Brad's performance has decreased** compared to the overall prompts graph, with a **drop of** (**-1.4%**).

On the other hand, **draws** has **increased** by **+1.3%**.

Finally, **ChatGPT** has **remained** the same.

The changes are barely perceptible so no conclusions can be drawn.

3.13.3 Overall Rewriting, (hyperspecific prompts)



For **hyperspecific prompts**, we can see how **Brad's performance has increased** compared to the **simple prompts** graph, with **+2.7%**.

Finally, **ChatGPT** has **remained** the same.

We can conclude that the **Bard model** has achieved a **higher number of successes with hyperspecific prompts**.

3.13.4 Conclusion Rewriting

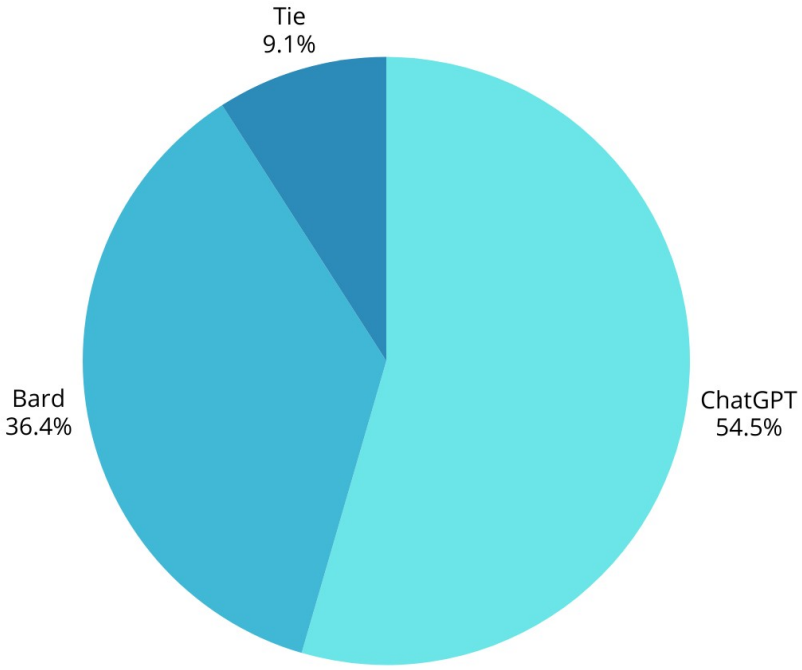
In conclusion, **ChatGPT is far superior in Rewriting**, with a score of **70.2%, while Bard had 20.3%.**

For **simple prompts**, **Brad's performance decreased** compared to the overall prompts graph, with a **drop of** (**-1.4%**).

For **hyperspecific prompts**, **Brad's performance increased** compared to the **simple prompts** graph, **+2.7%**.

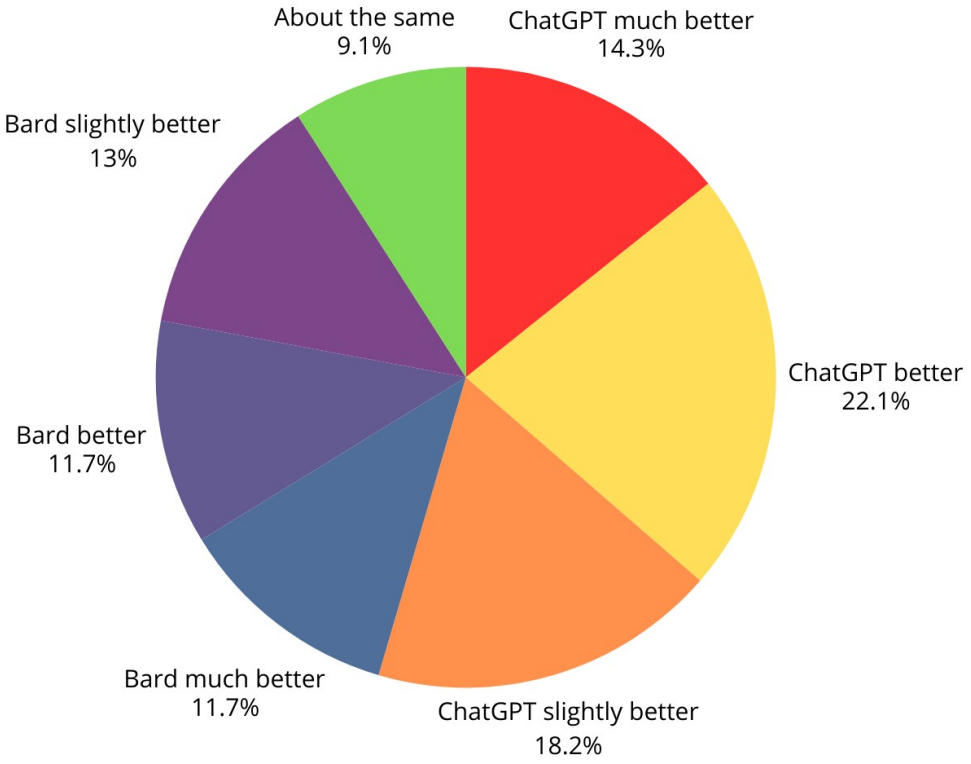
This means that for **hyperspecific prompts**, **Bard has been slightly better than with simple prompts**, even though ChatGPT is the clear winner.

3.14 Overall and Detailed Summarization Category



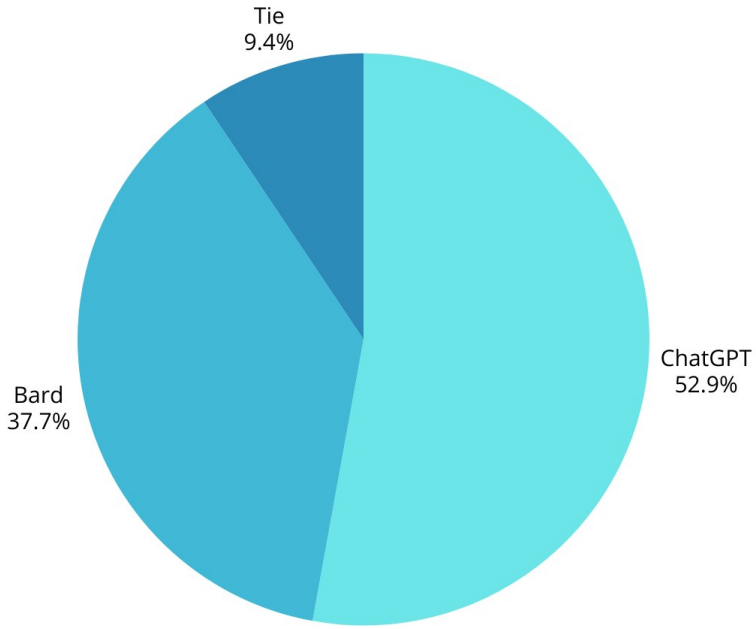
For this category, we can see that **ChatGPT's** answers were chosen by users **54.5%** of the time, while **Brad's answers** were chosen **36.4%** of the time.

There are hardly any ties so we can expect different answers.



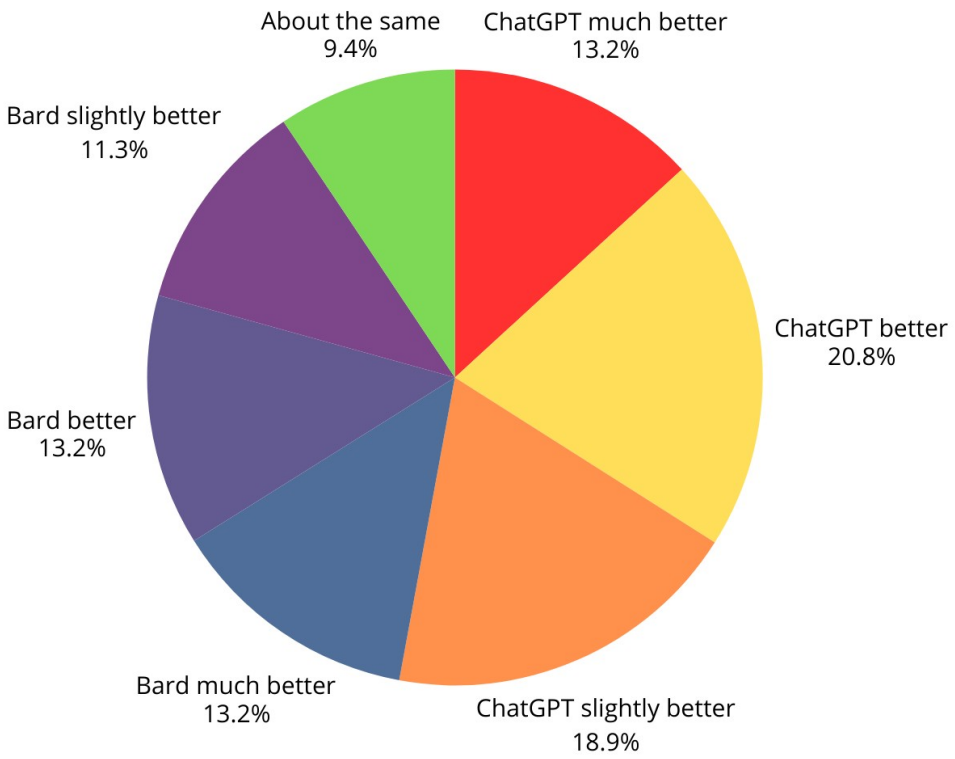
The data are similar except that **ChatGPT Better** is superior to **Bard Much Better**, with **22.1% vs 11.7%**

3.14.2 Overall and Detailed Summarization Category, (simple prompts)



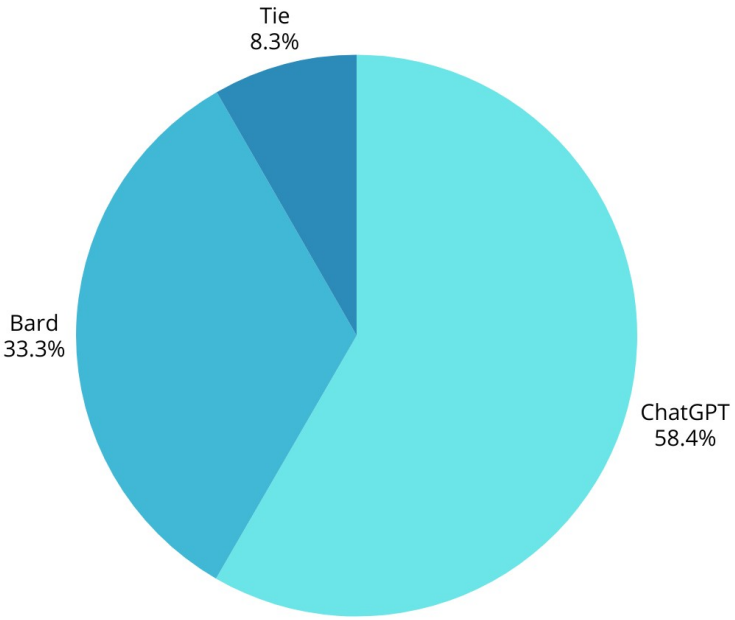
For **simple prompts**, we can see how **ChatGPT performance has decreased** compared to the overall prompts graph, with a **drop of** (**-1.6%**).

On the other hand, **Bard** has **increased** by **+1.3%**.



We can see how **ChatGPT Much Better** drops (**-1.1%**) while **Bard much better** rises **+1.5%**, leaving both at the same percentage of **13.2% Much Better.**

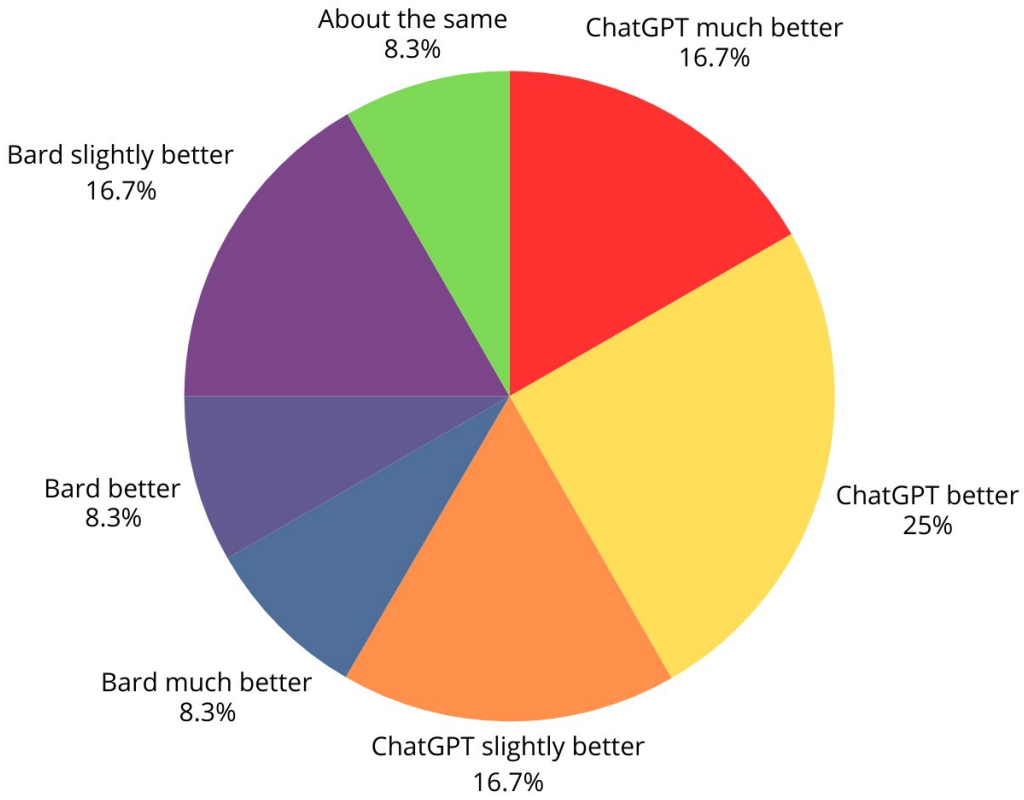
3.14.3 Overall and Detailed Summarization, (hyperspecific prompts)



For **hyperspecific prompts**, we can see how **ChatGPT performance has increased** compared to the **simple prompts** graph, **+5.5%**.

Bard **decreased** (**-4.4%**) compared to the **simple prompts.**

This means that Bard has had difficulties with hyperspecific prompts.



Moreover, if we compare the detailed graph of hyperspecific prompts with that of simple prompts we will see that:

Bard Much Better and Bard Better responses drop (-4.9%)

Between that, the fact that in general, compared to the **simple prompt Bard decreased (-4.4%),** and that **ChatGPT Much Better Response increases by +3.5%,** we can say that **Bard has had more difficulties with hyperspecific prompts.**

3.14.4 Conclusion Summarization

In conclusion, **ChatGPT is superior in Summarization**, with a score of **54.5%, while Bard had 36.4%.**

For **simple prompts**, we can see how **ChatGPT performance has decreased** compared to the overall prompts graph, with a **drop of** (**-1.6%**), while **Bard** has **increased** by **+1.3%**.

For **hyperspecific prompts**, **Brad's Much Better and Better responses decreased (-4.9%),** compared to the **simple prompts.**

Also for **hyperspecific prompts**, **ChatGPT Much Better Response increased by +3.5%** compared to the **simple prompt.**

So **Bard has had more difficulties with hyperspecific prompts than simple prompts.**

4. Analysis of explanations and language

Thanks to the Python code, we obtain the following information when searching for patterns and repeated words in the file:

~

Most common words in ChatGPT explanations (filtered):
[('better', 266), ('much', 143), ('prompt', 129), ('poem', 103), ('text', 93), ('provided', 92), ('correct', 71), ('code', 67), ('me', 66), ('request', 65)]

Most common words in Bard explanations (filtered):
[('better', 95), ('prompt', 33), ('much', 33), ('correct', 25), ('question', 25), ('provided', 24), ('included', 24), ('me', 23), ('slightly', 22), ('like', 22)]

Most common trigrams in ChatGPT explanations:
rated much better - 10
these reasons rated - 7
much better followed - 7
5 gallon jug - 5
3 gallon jug - 4
so much better - 4
m language model - 4
better much better - 4
so slightly better - 4
text based ai - 4

Most common trigrams in Bard explanations:
slightly better since - 4
jeff final count - 2
so much better - 2
type 1 diabetes - 2
models answered question - 2
angry text message - 2
does better job - 2
much better correct - 2
connection between golden - 2
between golden room - 2

Example ChatGPT explanations:
["While the supporting information given in Bard was good to know it wasn't requested. Both responses do well by giving a general rundown of the series, who

stars in it and its plot. ", 'ChatGPT does a lot more with the laptop-theme of the recipe. Bard creates a much more straight-forward recipe.', "ChatGPT's response is better because it follows the format of a haiku which is 3 lines with 5 syllables, 7 syllables, 5 syllables. The response by Bard follows the 3 lines rule but does not have the correct amount of syllables for each line. Both responses did a good job of following the requested topic."]

Example Bard explanations:
['Simple steps are all that are needed. Bard explained the answer simply, and it was correct. ', 'Bard has better formatting and writing queries I prefer the way Bard organises its information and its answering. The information is presented in a way I personally find more appealing.', 'Both effectively laid out the steps, and came to the correct answer, while explaining what a factorial is correct. They also both maintained the character pretty well. However, Bard managed to find a shortcut on repetitious steps by explaining that you just repeat that step with ever-decreasing numbers. This simplifies the output and is much more readable while still getting the point across.']

~

From there we can draw the following conclusions:

ChatGPT:

- It uses the word "better" more (266 vs. 95 in Bard), suggesting that his explanations consistently emphasize the superiority of one answer over another.
- It displays vocabulary more focused on the structure of prompts (prompt, poem, text, code), indicating that it tends to justify based on format and technical compliance.
- It most common trigrams ("rated much better," "these reasons rated," "much better followed") reflect a pattern of comparative and reasoned evaluation.
- It has a greater tendency to reason about formal accuracy: metrics in haikus, code structure, response format.

Bard:

- His explanations focus more on simplicity and readability (“simple steps,” “better formatting,” “much more readable”).

- Frequently used words like “question, included, slightly, like” suggest an approach more oriented toward the reader/user experience, rather than the structure of the prompt.
- His trigrams (“slightly better since,” “does better job,” “much better correct”) are more varied and less repetitive than ChatGPT's, which denotes a less rigid style.
- He adds value through clear and concise explanations, with a more subjective tone (“I prefer the way Bard organizes…”).

5. Performance evaluations

Thanks to the Python code, we obtain the following information when searching for patterns and repeated words in the file:

~

Performance/optimization comments for ChatGPT:
{'better': 266, 'correct': 71, 'detailed': 18, 'useful': 6, 'efficient': 1, 'accurate': 17, 'clear': 17, 'helpful': 20, 'quick': 1}

Performance/optimization comments for Bard:
{'better': 95, 'correct': 25, 'faster': 1, 'helpful': 16, 'accurate': 8, 'detailed': 4, 'speed': 2, 'useful': 3, 'quick': 2, 'clear': 2, 'improve': 1, 'fast': 1, 'efficient': 1}

~

From there we can draw the following conclusions:

ChatGPT received significantly more comments associated with "correct" (71 vs. 25), "detailed" (18 vs. 4), and "helpful" (20 vs. 16) → it is perceived as more complete, precise, and explanatory.

Bard stands out more for "faster / speed / quick" (several mentions) → it is associated with speed and efficiency, even if it is less detailed.

6. Sentiment in evaluations

Thanks to the Python code, we obtain the following information when searching for patterns and repeated words in the file:

~

Sentiment analysis in ChatGPT explanations:
{'positive': 347, 'negative': 30, 'neutral': 217}

Sentiment analysis in Bard explanations:
{'positive': 129, 'negative': 8, 'neutral': 109}

~

From there we can draw the following conclusions:

ChatGPT: More feedback (347 positive vs. 129 for Bard), but also more negative (30 vs. 8). → In other words, his answers are highly appreciated, but he also receives more criticism when he fails.

Bard: Fewer mentions overall, with a lower profile → Less enthusiastic but more consistent (fewer criticisms).

7. Errors pointed out

Thanks to the Python code, we obtain the following information when searching for patterns and repeated words in the file:

~

Errors directed at ChatGPT:
{'wrong': 1, 'incorrect': 1}
- Chat ChatGPT's response is better because it is formated as a text message per my request. It also does a better job of mimicking Obama's tone. The only place it fails is that it signs itself as from Obama, while I was only trying to mimic his tone not send the message from him.
- Both models included letters that weren't provided and hallucinated words that aren't in the dictionary. ChatGPT also included words that were less than 7 letters long, against the prompt's instructions.
- Both offer an adequate interpretation of the poem. ChatGPT does oddly get the number of stanzas wrong however, saying that there is 4 rather than 3.
- Both responses correctly classified the majority of the clothing items in my list but ChatGPT missed one. ChatGPT incorrectly classified "Corduroy" as an article of clothing when it is a type of fabric.
- ChatGPT failed to make the poem 2 quatrains long. They both rhymed pretty well but both failed to make the first and last words of the poem to be hippopotamus.

Errors directed at Bard:
{'incorrect': 1, 'hallucinate': 2, 'wrong': 2, 'error': 1}
- Both listings contain multiple errors and so I have them ranked as about equal. Bard even lists a source, but the information it provides doesn’t actually match up with the source.
- Based on Wikipedia, it looks like Bard is incorrect in stating that Ireland uses the term soccer to refer to Gaelic Football: "Association football (Irish: Sacar) commonly referred to as football or soccer, is the team sport with the highest level of participation in the Republic of Ireland (with five-a-side games being included).[1]It is the second most popular sport in Ireland[2][3] and it is also the third most popular spectator sport overall with 16% of total attendances at sports events, behind only Gaelic football (34%) and hurling (23%).[4] The national governing body for the sport is the Football Association of Ireland, which runs the national football team and the League of Ireland, which is the top level of the sport in the country. The term "football" is used interchangeably in Ireland between association football and Gaelic football. Rugby union, another popular type of football, is generally called "rugby", while rugby league, Australian Rules football and American football are niche and minority sports usually referred to by their long title."

- Bard did not give into the conspiracies or hallucinate information. It kept itself factual, denying my attempts, and instead gave the correct reasons on who built pyramids and why. It also stuck to a simplistic style of writing like I wanted.
- GPT solved the equation correctly and proved the root that it found. Bard not only solved the equation incorrectly but also tried to prove the wrong root as correct by manipulating the results to its advantage. 59/5-10 = 124/5-9/5 = -9/5
- Bard explains why it will not answer the question, choosing to offer more information and explaining the error in the user's way of thinking. That goes in line with the helpfulness and harmlessness policy that ensures no hate is spread.

~

From there we can draw the following conclusions:

ChatGPT: More formal errors (number of syllables, classifications, poem length). → These are minor technical errors, not so much content errors.

Bard: More serious errors (hallucinations, incorrect facts, manipulation of mathematical results, contradictions with sources). → Higher risk of factual or fabricated errors.

8. Final Conclusion

ChatGPT was the **best AI model**. It **was selected 59.2%** while **Bard was selected 24.5%**.
ChatGPT performed better in every category.

The categories ChatGPT performed by best in order are:

- Poetry: 83.5%
- Creative Writting: 73.7%
- Coding: 71.7%
- Rewriting: 70.2%
- Brainstorming: 67.1%
- Adversarial Dishonesty: 61.5%
- Extraction: 58.4%
- Summarization: 54.5%
- Classification: 49.3%
- Mathematical Reasoning: 48.8%
- Closed QA: 48.4%
- Open QA: 43%
- Adversarial Harmfulness: 41.4%

The categories Bard performed by best in order are:

- Summarization: 36.4%
- Open QA: 33.7%
- Mathematical Reasoning: 31.2%
- Closed QA: 30.1%
- Classification: 29%
- Adversarial Harmfulness: 25.7%
- Extraction: 22.1%
- Adversarial Dishonesty: 21.4%
- Coding: 20.8%
- Rewriting: 20.3%
- Brainstorming: 19.7%
- Creative Writting: 15.2%
- Poetry: 12.7%

As for **Bard**, when it comes to **responding to simple prompts** alone, his **success rate increased by +3.8%**.

So **Bard responds best to simple prompts**.

As for **Bard**, when it comes to **responding to hyperspecific prompts** alone, his **success rate decreased** by (**-8.3%**)

So **Bard responds worse to hyperspecific prompts**.

However, there are exceptions where in several categories it has responded better to hyperspecific prompts than to simple prompts.

- **Poetry: +8%**
- **Adversial Dishonesty: +3.6%**
- **Rewritting: +1.3%**

And finally, based on the **Analysis of explanations and language**, **Performance evaluations**, **Sentiment in evaluations** and **Errors pointed out**, we can reach this conclusion:

ChatGPT stands out for:

- Longer and more justified explanations.
- Focus on technical correctness, formatting, and accuracy.
- Perceived as more "academic/professional," although it can be overly verbose and make minor errors of detail.

Bard stands out for:

- Simpler, more legible, and faster answers.
- Good for tasks where clarity is more important than technical precision.
- Risk: May provide incorrect or fabricated information more frequently.