Artificial Phantasia: Evidence for Propositional Reasoning-Based Mental Imagery Within Large Language Models

Author

August 28, 2025

Summary

This R Markdown document reproduces the analyses reported in [Author Names Removed for Anonymized Peer Review]. *Artificial Phantasia: Evidence for Propositional Reasoning-Based Mental Imagery Within Large Language Models.*

`Groundhog` will load the exact versions of the `R` packages used for the reported analyses. However, it cannot control the version of `R` that you are running. We used `\R 4.5.1`. `Groundhog` will load `tidyverse 2.0.0`, `knitr 1.50`, and `patchwork 1.3.1`.

If you have issues with `groundhog` (authorizing in the console the creation of a library folder when using groundhog for the first time is needed) or do not want to use it, follow the instructions in the comment below.

```
llm_data_finke <- read.csv("output_csvs/llm_graded_results_finke.csv")
llm_data_novel <- read.csv("output_csvs/llm_graded_results_novel.csv")

human_data_finke <- read.csv("output_csvs/h_graded_results_finke.csv")
human_data_novel <- read.csv("output_csvs/h_graded_results_novel.csv")
```

```
# Data
## Finke et al. Tasks
humans_finke_score <- sum(human_data_finke$overall_score)
humans_finke_max_score <- sum(human_data_finke$n_total) * 5

o3_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI o3 Single Context (2025-07-21)", "overa
  llm_data_finke[llm_data_finke$Model == "OpenAI o3 Multiple Context (2025-07-21)", "overall_score"]
o3_finke_max_score <- (llm_data_finke[llm_data_finke$Model == "OpenAI o3 Single Context (2025-07-21)",
  llm_data_finke[llm_data_finke$Model == "OpenAI o3 Multiple Context (2025-07-21)", "n_total"]) * 5

o3_images_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI o3 Multiple Context w/ Images (
o3_images_finke_max_score <- llm_data_finke[llm_data_finke$Model == "OpenAI o3 Multiple Context w/ Image

o3_pro_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI o3 Pro Single Context (2025-07-21)
  llm_data_finke[llm_data_finke$Model == "OpenAI o3 Pro Multiple Context (2025-07-21)", "overall_score"
o3_pro_finke_max_score <- (llm_data_finke[llm_data_finke$Model == "OpenAI o3 Pro Single Context (2025-0
  llm_data_finke[llm_data_finke$Model == "OpenAI o3 Pro Multiple Context (2025-07-21)", "n_total"]) * 5

o4_mini_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI o4-mini Multiple Context (2025-07-
  llm_data_finke[llm_data_finke$Model == "OpenAI o4-mini Single Context (2025-07-14)", "overall_score"]
  llm_data_finke[llm_data_finke$Model == "OpenAI o4-mini Multiple Context (2025-07-21)", "overall_score
  llm_data_finke[llm_data_finke$Model == "OpenAI o4-mini Single Context (2025-07-21)", "overall_score"]
o4_mini_finke_max_score <- (llm_data_finke[llm_data_finke$Model == "OpenAI o4-mini Multiple Context (20
```

```r
  llm_data_finke[llm_data_finke$Model == "OpenAI o4-mini Single Context (2025-07-14)", "n_total"] +
  llm_data_finke[llm_data_finke$Model == "OpenAI o4-mini Multiple Context (2025-07-21)", "n_total"] +
  llm_data_finke[llm_data_finke$Model == "OpenAI o4-mini Single Context (2025-07-21)", "n_total"]) * 5

chatgpt_4o_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI ChatGPT-4o Single Context (2025
  llm_data_finke[llm_data_finke$Model == "OpenAI ChatGPT-4o Multiple Context (2025-07-25)", "overall_sco
chatgpt_4o_finke_max_score <- (llm_data_finke[llm_data_finke$Model == "OpenAI ChatGPT-4o Single Context
  llm_data_finke[llm_data_finke$Model == "OpenAI ChatGPT-4o Multiple Context (2025-07-25)", "n_total"])

gpt4_1_finke_score <- llm_data_finke[llm_data_finke$Model == "GPT 4.1 Single Context (2025-07-21)", "ove
  llm_data_finke[llm_data_finke$Model == "GPT 4.1 Multiple Context (2025-07-21)", "overall_score"]
gpt4_1_finke_max_score <- (llm_data_finke[llm_data_finke$Model == "GPT 4.1 Single Context (2025-07-21)"
  llm_data_finke[llm_data_finke$Model == "GPT 4.1 Multiple Context (2025-07-21)", "n_total"]) * 5

gpt4_1_images_finke_score <- llm_data_finke[llm_data_finke$Model == "GPT 4.1 Multiple Context w/ Images
  llm_data_finke[llm_data_finke$Model == "GPT 4.1 Single Context w/ Images (2025-07-21)", "overall_scor
gpt4_1_images_finke_max_score <- (llm_data_finke[llm_data_finke$Model == "GPT 4.1 Multiple Context w/ I
  llm_data_finke[llm_data_finke$Model == "GPT 4.1 Single Context w/ Images (2025-07-21)", "n_total"]) *

gemini2_5_finke_score <- llm_data_finke[llm_data_finke$Model == "Gemini 2.5 Pro Multiple Context (2025-
  llm_data_finke[llm_data_finke$Model == "Gemini 2.5 Pro Single Context (2025-07-21)", "overall_score"]
gemini2_5_finke_max_score <- (llm_data_finke[llm_data_finke$Model == "Gemini 2.5 Pro Multiple Context (
  llm_data_finke[llm_data_finke$Model == "Gemini 2.5 Pro Single Context (2025-07-21)", "n_total"]) * 5

gemini2_0_flash_finke_score <- llm_data_finke[llm_data_finke$Model == "Gemini 2.0 Flash Multiple Contex
  llm_data_finke[llm_data_finke$Model == "Gemini 2.0 Flash Single Context (2025-07-21)", "overall_score
gemini2_0_flash_finke_max_score <- (llm_data_finke[llm_data_finke$Model == "Gemini 2.0 Flash Multiple C
  llm_data_finke[llm_data_finke$Model == "Gemini 2.0 Flash Single Context (2025-07-21)", "n_total"]) *

gemini2_0_flash_images_finke_score <- llm_data_finke[llm_data_finke$Model == "Gemini 2.0 Flash Image Ge
gemini2_0_flash_images_finke_max_score <- llm_data_finke[llm_data_finke$Model == "Gemini 2.0 Flash Imag
```

```r
## Novel 48 Tasks
humans_novel_score <- sum(human_data_novel$overall_score)
humans_novel_max_score <- sum(human_data_novel$n_total) * 5

o3_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI o3 Single Context (2025-07-21)", "overa
  llm_data_novel[llm_data_novel$Model == "OpenAI o3 Multiple Context (2025-07-21)", "overall_score"]
o3_novel_max_score <- (llm_data_novel[llm_data_novel$Model == "OpenAI o3 Single Context (2025-07-21)",
  llm_data_novel[llm_data_novel$Model == "OpenAI o3 Multiple Context (2025-07-21)", "n_total"]) * 5

o3_images_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI o3 Multiple Context w/ Images (
o3_images_novel_max_score <- llm_data_novel[llm_data_novel$Model == "OpenAI o3 Multiple Context w/ Image

o3_pro_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI o3 Pro Single Context (2025-07-21)
  llm_data_novel[llm_data_novel$Model == "OpenAI o3 Pro Multiple Context (2025-07-21)", "overall_score"
o3_pro_novel_max_score <- (llm_data_novel[llm_data_novel$Model == "OpenAI o3 Pro Single Context (2025-0
  llm_data_novel[llm_data_novel$Model == "OpenAI o3 Pro Multiple Context (2025-07-21)", "n_total"]) * 5

o4_mini_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI o4-mini Multiple Context (2025-07
  llm_data_novel[llm_data_novel$Model == "OpenAI o4-mini Single Context (2025-07-14)", "overall_score"]
  llm_data_novel[llm_data_novel$Model == "OpenAI o4-mini Multiple Context (2025-07-21)", "overall_score
  llm_data_novel[llm_data_novel$Model == "OpenAI o4-mini Single Context (2025-07-21)", "overall_score"]
o4_mini_novel_max_score <- (llm_data_novel[llm_data_novel$Model == "OpenAI o4-mini Multiple Context (20
  llm_data_novel[llm_data_novel$Model == "OpenAI o4-mini Single Context (2025-07-14)", "n_total"] +
  llm_data_novel[llm_data_novel$Model == "OpenAI o4-mini Multiple Context (2025-07-21)", "n_total"] +
  llm_data_novel[llm_data_novel$Model == "OpenAI o4-mini Single Context (2025-07-21)", "n_total"]) * 5

chatgpt_4o_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI ChatGPT-4o Single Context (2025
  llm_data_novel[llm_data_novel$Model == "OpenAI ChatGPT-4o Multiple Context (2025-07-25)", "overall_sc
chatgpt_4o_novel_max_score <- (llm_data_novel[llm_data_novel$Model == "OpenAI ChatGPT-4o Single Context
  llm_data_novel[llm_data_novel$Model == "OpenAI ChatGPT-4o Multiple Context (2025-07-25)", "n_total"])

gpt4_1_novel_score <- llm_data_novel[llm_data_novel$Model == "GPT 4.1 Single Context (2025-07-21)", "ove
  llm_data_novel[llm_data_novel$Model == "GPT 4.1 Multiple Context (2025-07-21)", "overall_score"]
gpt4_1_novel_max_score <- (llm_data_novel[llm_data_novel$Model == "GPT 4.1 Single Context (2025-07-21)"
  llm_data_novel[llm_data_novel$Model == "GPT 4.1 Multiple Context (2025-07-21)", "n_total"]) * 5

gpt4_1_images_novel_score <- llm_data_novel[llm_data_novel$Model == "GPT 4.1 Multiple Context w/ Images
  llm_data_novel[llm_data_novel$Model == "GPT 4.1 Single Context w/ Images (2025-07-21)", "overall_scor
gpt4_1_images_novel_max_score <- (llm_data_novel[llm_data_novel$Model == "GPT 4.1 Multiple Context w/ In
  llm_data_novel[llm_data_novel$Model == "GPT 4.1 Single Context w/ Images (2025-07-21)", "n_total"]) *

gemini2_5_novel_score <- llm_data_novel[llm_data_novel$Model == "Gemini 2.5 Pro Multiple Context (2025-(
  llm_data_novel[llm_data_novel$Model == "Gemini 2.5 Pro Single Context (2025-07-21)", "overall_score"]
gemini2_5_novel_max_score <- (llm_data_novel[llm_data_novel$Model == "Gemini 2.5 Pro Multiple Context (2
  llm_data_novel[llm_data_novel$Model == "Gemini 2.5 Pro Single Context (2025-07-21)", "n_total"]) * 5

gemini2_0_flash_novel_score <- llm_data_novel[llm_data_novel$Model == "Gemini 2.0 Flash Multiple Context
  llm_data_novel[llm_data_novel$Model == "Gemini 2.0 Flash Single Context (2025-07-21)", "overall_score
gemini2_0_flash_novel_max_score <- (llm_data_novel[llm_data_novel$Model == "Gemini 2.0 Flash Multiple Co
  llm_data_novel[llm_data_novel$Model == "Gemini 2.0 Flash Single Context (2025-07-21)", "n_total"]) * 5

gemini2_0_flash_images_novel_score <- llm_data_novel[llm_data_novel$Model == "Gemini 2.0 Flash Image Ger
gemini2_0_flash_images_novel_max_score <- llm_data_novel[llm_data_novel$Model == "Gemini 2.0 Flash Image
```

```r
## Collapse Families
o3_family_finke_score <- o3_finke_score +
  o3_images_finke_score +
  o3_pro_finke_score
o3_family_finke_max_score <- o3_finke_max_score +
  o3_images_finke_max_score +
  o3_pro_finke_max_score

o3_family_novel_score <- o3_novel_score +
  o3_images_novel_score +
  o3_pro_novel_score
o3_family_novel_max_score <- o3_novel_max_score +
  o3_images_novel_max_score +
  o3_pro_novel_max_score


o3_family_no_images_finke_score <- o3_finke_score + o3_pro_finke_score
o3_family_no_images_finke_max_score <- o3_finke_max_score + o3_pro_finke_max_score

o3_family_no_images_novel_score <- o3_novel_score + o3_pro_novel_score
o3_family_no_images_novel_max_score <- o3_novel_max_score + o3_pro_novel_max_score
```

```r
gemini_family_finke_score <- gemini2_5_finke_score +
  gemini2_0_flash_finke_score +
  gemini2_0_flash_images_finke_score
gemini_family_finke_max_score <- gemini2_5_finke_max_score +
  gemini2_0_flash_finke_max_score +
  gemini2_0_flash_images_finke_max_score

gemini_family_novel_score <- gemini2_5_novel_score +
  gemini2_0_flash_novel_score +
  gemini2_0_flash_images_novel_score
gemini_family_novel_max_score <- gemini2_5_novel_max_score +
  gemini2_0_flash_novel_max_score +
  gemini2_0_flash_images_novel_max_score

openai_family_finke_score <- o3_family_finke_score +
  o4_mini_finke_score +
  chatgpt_4o_finke_score +
  gpt4_1_finke_score +
  gpt4_1_images_finke_score
openai_family_finke_max_score <- o3_family_finke_max_score +
  o4_mini_finke_max_score +
  chatgpt_4o_finke_max_score +
  gpt4_1_finke_max_score +
  gpt4_1_images_finke_max_score

openai_family_novel_score <- o3_family_novel_score +
  o4_mini_novel_score +
  chatgpt_4o_novel_score +
  gpt4_1_novel_score +
  gpt4_1_images_novel_score
openai_family_novel_max_score <- o3_family_novel_max_score +
  o4_mini_novel_max_score +
  chatgpt_4o_novel_max_score +
  gpt4_1_novel_max_score +
  gpt4_1_images_novel_max_score

openai_non_o3_family_finke_score <- o4_mini_finke_score +
  chatgpt_4o_finke_score +
  gpt4_1_finke_score +
  gpt4_1_images_finke_score
openai_non_o3_family_finke_max_score <- o4_mini_finke_max_score +
  chatgpt_4o_finke_max_score +
  gpt4_1_finke_max_score +
  gpt4_1_images_finke_max_score

openai_non_o3_family_novel_score <- o4_mini_novel_score +
  chatgpt_4o_novel_score +
  gpt4_1_novel_score +
  gpt4_1_images_novel_score
openai_non_o3_family_novel_max_score <- o4_mini_novel_max_score +
  chatgpt_4o_novel_max_score +
  gpt4_1_novel_max_score +
  gpt4_1_images_novel_max_score
```

```
## Collapsed Data (Finke + 48 Novel)
humans_total_score <- humans_finke_score + humans_novel_score
humans_total_max_score <- humans_finke_max_score + humans_novel_max_score

o3_total_score <- o3_finke_score + o3_novel_score
o3_total_max_score <- o3_finke_max_score + o3_novel_max_score

o3_images_total_score <- o3_images_finke_score + o3_images_novel_score
o3_images_total_max_score <- o3_images_finke_max_score + o3_images_novel_max_score

o3_pro_total_score <- o3_pro_finke_score + o3_pro_novel_score
o3_pro_total_max_score <- o3_pro_finke_max_score + o3_pro_novel_max_score

o4_mini_total_score <- o4_mini_finke_score + o4_mini_novel_score
o4_mini_total_max_score <- o4_mini_finke_max_score + o4_mini_novel_max_score

chatgpt_4o_total_score <- chatgpt_4o_finke_score + chatgpt_4o_novel_score
chatgpt_4o_total_max_score <- chatgpt_4o_finke_max_score + chatgpt_4o_novel_max_score

gpt4_1_total_score <- gpt4_1_finke_score + gpt4_1_novel_score
gpt4_1_total_max_score <- gpt4_1_finke_max_score + gpt4_1_novel_max_score

gpt4_1_images_total_score <- gpt4_1_images_finke_score + gpt4_1_images_novel_score
gpt4_1_images_total_max_score <- gpt4_1_images_finke_max_score + gpt4_1_images_novel_max_score

gemini2_5_total_score <- gemini2_5_finke_score + gemini2_5_novel_score
gemini2_5_total_max_score <- gemini2_5_finke_max_score + gemini2_5_novel_max_score

gemini2_0_flash_total_score <- gemini2_0_flash_finke_score + gemini2_0_flash_novel_score
gemini2_0_flash_total_max_score <- gemini2_0_flash_finke_max_score + gemini2_0_flash_novel_max_score

gemini2_0_flash_images_total_score <- gemini2_0_flash_images_finke_score + gemini2_0_flash_images_novel_
gemini2_0_flash_images_total_max_score <- gemini2_0_flash_images_finke_max_score + gemini2_0_flash_imag

o3_family_total_score <- o3_family_finke_score + o3_family_novel_score
o3_family_total_max_score <- o3_family_finke_max_score + o3_family_novel_max_score

o3_family_no_images_total_score <- o3_family_no_images_finke_score + o3_family_no_images_novel_score
o3_family_no_images_total_max_score <- o3_family_no_images_finke_max_score + o3_family_no_images_novel_

all_gemini_total_score <- gemini_family_finke_score + gemini_family_novel_score
all_gemini_total_max_score <- gemini_family_finke_max_score + gemini_family_novel_max_score

all_openAI_total_score <- openai_family_finke_score + openai_family_novel_score
all_openAI_total_max_score <- openai_family_finke_max_score + openai_family_novel_max_score

other_openAI_total_score <- openai_non_o3_family_finke_score + openai_non_o3_family_novel_score
other_openAI_total_max_score <- openai_non_o3_family_finke_max_score + openai_non_o3_family_novel_max_s

## Original Finke Data - modified towards the new scoring system
original_finke_exp2_correct <- 37 * 5 + 72 - 37
original_finke_exp2_total <- 72 * 5
```

```r
original_finke_exp3_correct <- 28 * 5 + 72 - 28
original_finke_exp3_total <- 72 * 5

# Collapsed Original Finke (Exp 2 + Exp 3)
original_finke_correct <- original_finke_exp2_correct + original_finke_exp3_correct
original_finke_total <- original_finke_exp2_total + original_finke_exp3_total
```

```r
# Create data frames for easier manipulation
finke_data <- data.frame(
  model = c("Humans", "o3", "o3-GPT-Image",
            "o3-Pro", "GPT-4.1", "GPT-4.1-GPT-Image",
            "ChatGPT-4o", "o4-mini", "Gemini-2.5",
            "Gemini-2.0-Flash", "Gemini-2.0-Flash-GPT-Image",
            "All-OpenAI", "Other-OpenAI", "All-Gemini"),
  correct = c(humans_finke_score, o3_finke_score, o3_images_finke_score,
              o3_pro_finke_score, gpt4_1_finke_score, gpt4_1_images_finke_score,
              chatgpt_4o_finke_score, o4_mini_finke_score, gemini2_5_finke_score,
              gemini2_0_flash_finke_score, gemini2_0_flash_images_finke_score,
              openai_family_finke_score, openai_non_o3_family_finke_score, gemini_family_finke_score),
  total = c(humans_finke_max_score, o3_finke_max_score, o3_images_finke_max_score,
            o3_pro_finke_max_score, gpt4_1_finke_max_score, gpt4_1_images_finke_max_score,
            chatgpt_4o_finke_max_score, o4_mini_finke_max_score, gemini2_5_finke_max_score,
            gemini2_0_flash_finke_max_score, gemini2_0_flash_images_finke_max_score,
            openai_family_finke_max_score, openai_non_o3_family_finke_max_score, gemini_family_finke_ma
)

# Calculate proportions from correct/total
finke_data$proportion <- finke_data$correct / finke_data$total
```

```r
novel_data <- data.frame(
  model = c("Humans", "o3", "o3-GPT-Image",
            "o3-Pro", "GPT-4.1", "GPT-4.1-GPT-Image",
            "ChatGPT-4o", "o4-mini", "Gemini-2.5",
            "Gemini-2.0-Flash", "Gemini-2.0-Flash-GPT-Image",
            "All-OpenAI", "Other-OpenAI", "All-Gemini"),
    correct = c(humans_novel_score, o3_novel_score, o3_images_novel_score,
                o3_pro_novel_score, gpt4_1_novel_score, gpt4_1_images_novel_score,
                chatgpt_4o_novel_score, o4_mini_novel_score, gemini2_5_novel_score,
                gemini2_0_flash_novel_score, gemini2_0_flash_images_novel_score,
                openai_family_novel_score, openai_non_o3_family_novel_score, gemini_family_novel_score)
    total = c(humans_novel_max_score, o3_novel_max_score, o3_images_novel_max_score,
                o3_pro_novel_max_score, gpt4_1_novel_max_score, gpt4_1_images_novel_max_score,
                chatgpt_4o_novel_max_score, o4_mini_novel_max_score, gemini2_5_novel_max_score,
                gemini2_0_flash_novel_max_score, gemini2_0_flash_images_novel_max_score,
                openai_family_novel_max_score, openai_non_o3_family_novel_max_score, gemini_family_novel
)

# Calculate proportions from correct/total
novel_data$proportion <- novel_data$correct / novel_data$total
```

```r
collapsed_data <- data.frame(
    model = c("Humans", "o3", "o3-GPT-Image",
                "o3-Pro", "GPT-4.1", "GPT-4.1-GPT-Image",
                "ChatGPT-4o", "o4-mini", "Gemini-2.5",
                "Gemini-2.0-Flash", "Gemini-2.0-Flash-GPT-Image",
                "All-OpenAI", "Other-OpenAI", "All-Gemini"),
    correct = c(humans_total_score, o3_total_score, o3_images_total_score,
                o3_pro_total_score, gpt4_1_total_score, gpt4_1_images_total_score,
                chatgpt_4o_total_score, o4_mini_total_score, gemini2_5_total_score,
                gemini2_0_flash_total_score, gemini2_0_flash_images_total_score,
                all_openAI_total_score, other_openAI_total_score, all_gemini_total_score),
    total = c(humans_total_max_score, o3_total_max_score, o3_images_total_max_score,
                o3_pro_total_max_score, gpt4_1_total_max_score, gpt4_1_images_total_max_score,
                chatgpt_4o_total_max_score, o4_mini_total_max_score, gemini2_5_total_max_score,
                gemini2_0_flash_total_max_score, gemini2_0_flash_images_total_max_score,
                all_openAI_total_max_score, other_openAI_total_max_score, all_gemini_total_max_score)
    )



# Calculate proportions from correct/total
collapsed_data$proportion <- collapsed_data$correct / collapsed_data$total
```

```r
# Display the data
cat("Finke et al. Tasks Data:\n")
```

```
## Finke et al. Tasks Data:
```

```r
print(finke_data)
```

```
##                         model   correct total proportion
## 1                      Humans 953.75952  1525  0.6254161
## 2                          o3  77.55000   120  0.6462500
## 3                o3-GPT-Image  31.32143    60  0.5220238
## 4                      o3-Pro  80.70000   120  0.6725000
## 5                     GPT-4.1  56.25476   120  0.4687897
## 6           GPT-4.1-GPT-Image  40.25000   120  0.3354167
## 7                  ChatGPT-4o  49.36905   120  0.4114087
## 8                     o4-mini 113.63333   240  0.4734722
## 9                  Gemini-2.5  59.50000   120  0.4958333
## 10           Gemini-2.0-Flash  43.05952   120  0.3588294
## 11 Gemini-2.0-Flash-GPT-Image  20.90476    60  0.3484127
## 12                 All-OpenAI 449.07857   900  0.4989762
## 13               Other-OpenAI 259.50714   600  0.4325119
## 14                 All-Gemini 123.46429   300  0.4115476
```

```r
cat("\n48 Novel Tasks Data:\n")
```

```
##
## 48 Novel Tasks Data:
```

```r
print(novel_data)
```

```
##                         model    correct total proportion
## 1                      Humans 3085.39286  5965  0.5172494
## 2                          o3  283.44286   480  0.5905060
## 3                o3-GPT-Image  130.92143   240  0.5455060
## 4                      o3-Pro  301.23810   480  0.6275794
## 5                     GPT-4.1  193.84524   480  0.4038442
## 6           GPT-4.1-GPT-Image  190.62143   480  0.3971280
## 7                 ChatGPT-4o  200.38333   480  0.4174653
## 8                    o4-mini  481.31190   960  0.5013666
## 9                 Gemini-2.5  213.76905   480  0.4453522
## 10          Gemini-2.0-Flash  189.14524   480  0.3940526
## 11 Gemini-2.0-Flash-GPT-Image   72.44048   240  0.3018353
## 12                All-OpenAI 1781.76429  3600  0.4949345
## 13              Other-OpenAI 1066.16190  2400  0.4442341
## 14                All-Gemini  475.35476  1200  0.3961290
```

```
cat("\nCollapsed Data (Finke + 48 Novel Tasks):\n")
```

```
##
## Collapsed Data (Finke + 48 Novel Tasks):
```

```
print(collapsed_data)
```

```
##                         model    correct total proportion
## 1                      Humans 4039.15238  7490  0.5392727
## 2                          o3  360.99286   600  0.6016548
## 3                o3-GPT-Image  162.24286   300  0.5408095
## 4                      o3-Pro  381.93810   600  0.6365635
## 5                     GPT-4.1  250.10000   600  0.4168333
## 6           GPT-4.1-GPT-Image  230.87143   600  0.3847857
## 7                  ChatGPT-4o  249.75238   600  0.4162540
## 8                     o4-mini  594.94524  1200  0.4957877
## 9                  Gemini-2.5  273.26905   600  0.4554484
## 10           Gemini-2.0-Flash  232.20476   600  0.3870079
## 11 Gemini-2.0-Flash-GPT-Image   93.34524   300  0.3111508
## 12                 All-OpenAI 2230.84286  4500  0.4957429
## 13               Other-OpenAI 1325.66905  3000  0.4418897
## 14                 All-Gemini  598.81905  1500  0.3992127
```

```
# Display Original Finke data
cat("\n\nOriginal Finke Data:\n")
```

```
##
##
## Original Finke Data:
```

```
cat("Exp 2: ", original_finke_exp2_correct, "/", original_finke_exp2_total,
    " (", round(original_finke_exp2_correct / original_finke_exp2_total, 3), ")\n", sep = "")
```

```
## Exp 2: 220/360 (0.611)
```

```
cat("Exp 3: ", original_finke_exp3_correct, "/", original_finke_exp3_total,
    " (", round(original_finke_exp3_correct / original_finke_exp3_total, 3), ")\n", sep = "")
```

```
## Exp 3: 184/360 (0.511)
```

```
cat("Collapsed Original Finke: ", original_finke_correct, "/", original_finke_total,
    " (", round(original_finke_correct / original_finke_total, 3), ")\n", sep = "")
```

```
## Collapsed Original Finke: 404/720 (0.561)
```

Proportion Testing Function

```
# Function to perform proportion test and extract results
perform_prop_test <- function(model1_name, model1_correct, model1_total,
```

```r
                              model2_name, model2_correct, model2_total) {

  # Perform the test
  test_result <- prop.test(x = c(model1_correct, model2_correct),
                           n = c(model1_total, model2_total),
                           alternative = "two.sided",
                           conf.level = 0.95,
                           correct = TRUE)

  # Calculate proportions
  prop1 <- model1_correct / model1_total
  prop2 <- model2_correct / model2_total
  diff <- prop1 - prop2

  # Return results as a list
  return(list(
    comparison = paste(model1_name, "vs", model2_name),
    model1 = model1_name,
    model2 = model2_name,
    prop1 = prop1,
    prop2 = prop2,
    diff = diff,
    chi_squared = test_result$statistic,
    df = test_result$parameter,
    p_value = test_result$p.value,
    ci_lower = test_result$conf.int[1],
    ci_upper = test_result$conf.int[2],
    significant = test_result$p.value < 0.05
  ))
}

# Function to test all combinations
test_all_combinations <- function(data, task_name) {
  results <- list()
  counter <- 1

  # Test all unique pairs
  for (i in 1:(nrow(data) - 1)) {
    for (j in (i + 1):nrow(data)) {
      results[[counter]] <- perform_prop_test(
        data$model[i], data$correct[i], data$total[i],
        data$model[j], data$correct[j], data$total[j]
      )
      counter <- counter + 1
    }
  }

  # Convert to data frame
  results_df <- do.call(rbind, lapply(results, as.data.frame))
  results_df$task <- task_name

  return(results_df)
}
```

Comparison: o3 Family Single Context vs Multiple Context

```
##
##
## Comparison: o3 Family Single Context vs Multiple Context (Finke Tasks)
## ======================================================================
## o3 Single Context: 77.55/120 (0.646)
## o3 Multiple Context: 31.32143/60 (0.522)
## Difference:  0.124
## Chi-squared:  2.089
## P-value:  0.1484
## 95% CI: [ -0.041 ,  0.289 ]
## Significant:  NO
##
##
## Detailed Comparison: o3 Single Context vs Multiple Context
## ----------------------------------------
## Proportions:  0.646  vs  0.522
## Difference:  0.124
## Chi-squared:  2.089
## Degrees of freedom:  1
## P-value:  0.1484
## 95% CI: [ -0.041 ,  0.289 ]
## Significant:  NO
##
##
## Summary Table - o3 Single vs Multiple Context:
##
##
## comparison                                  diff   p_value  significant
## --------------------------------------  ------  --------  ------------
## o3 Single Context vs Multiple Context    0.124    0.1484  FALSE
```

Comparison: Current Human Finke vs Original Finke

```
##
##
## Comparison: Current Human Finke vs Original Finke (Collapsed Exp 2 + Exp 3)
## ======================================================================
## Current Human Finke: 953.7595/1525 (0.625)
## Original Finke: 404/720 (0.561)
## Difference:  0.064
## Chi-squared:  8.195
## P-value:  0.004202
## 95% CI: [ 0.02 ,  0.109 ]
## Significant:  YES (p < 0.05)
##
##
## Detailed Comparison: Current Humans vs Original Finke
## ----------------------------------------
## Proportions:  0.625  vs  0.561
## Difference:  0.064
## Chi-squared:  8.195
## Degrees of freedom:  1
## P-value:  0.004202
## 95% CI: [ 0.02 ,  0.109 ]
## Significant:  YES (p < 0.05)
##
##
## Summary Table - Human vs Original Finke:
##
##
## comparison                           diff   p_value  significant
## ---------------------------------  ------  --------  ------------
## Current Humans vs Original Finke    0.064    0.0042  TRUE
```

Comparison: Current Human 48 vs Original Finke

```
##
##
## Comparison: Current Human 48-Item Task vs Original Finke (Collapsed Exp 2 + Exp 3)
## ======================================================================
## Current Human 48: 3085.393/5965 (0.517)
## Original Finke: 404/720 (0.561)
## Difference:  -0.044
## Chi-squared:  4.779
## P-value:  0.0288
## 95% CI: [ -0.083 ,  -0.005 ]
## Significant:  YES (p < 0.05)
##
##
## Detailed Comparison: Current Humans vs Original Finke
## ----------------------------------------
## Proportions:  0.517  vs  0.561
## Difference:  -0.044
## Chi-squared:  4.779
## Degrees of freedom:  1
## P-value:  0.0288
## 95% CI: [ -0.083 ,  -0.005 ]
## Significant:  YES (p < 0.05)
##
##
## Summary Table - Human vs Original Finke:
##
##
## comparison                               diff   p_value  significant
## --------------------------------  -------  --------  ------------
## Current Humans vs Original Finke    -0.044    0.0288  TRUE
```

Comparison: Current Humans (collapsed) vs Original Finke

```
##
##
## Comparison: Current Human 48-Item Task vs Original Finke (Collapsed Exp 2 + Exp 3)
## ======================================================================
## Current Human Finke: 4039.152/7490 (0.539)
## Original Finke: 404/720 (0.561)
## Difference:  -0.022
## Chi-squared:  1.175
## P-value:  0.2783
## 95% CI: [ -0.061 ,  0.017 ]
## Significant:  NO
##
##
## Detailed Comparison: Current Humans vs Original Finke
## ----------------------------------------
## Proportions:  0.539  vs  0.561
## Difference:  -0.022
## Chi-squared:  1.175
## Degrees of freedom:  1
## P-value:  0.2783
## 95% CI: [ -0.061 ,  0.017 ]
## Significant:  NO
##
##
## Summary Table - Current Human (Collapsed) vs Original Finke:
```

```
##
##
## comparison                                           diff   p_value  significant
## ---------------------------------------------  -------  --------  ------------
## Current Humans (collapsed) vs Original Finke    -0.022    0.2783  FALSE
```

Finke et al. Tasks - All Pairwise Comparisons

```
# Test all combinations for Finke tasks
finke_results <- test_all_combinations(finke_data, "Finke")

# Display results
cat("All Pairwise Comparisons for Finke et al. Tasks:\n")
```

```
## All Pairwise Comparisons for Finke et al. Tasks:
```

```
cat(paste(rep("=", 80), collapse = ""), "\n")
```

```
## ================================================================================
```

```
for (i in 1:nrow(finke_results)) {
  cat("\n", finke_results$comparison[i], "\n")
  cat(paste(rep("-", 40), collapse = ""), "\n")
  cat("Proportions: ", round(finke_results$prop1[i], 3), " vs ",
      round(finke_results$prop2[i], 3), "\n")
  cat("Difference: ", round(finke_results$diff[i], 3), "\n")
  cat("Chi-squared: ", round(finke_results$chi_squared[i], 3), "\n")
  cat("Degrees of freedom: ", round(finke_results$df[i], 3), "\n")
  cat("P-value: ", format(finke_results$p_value[i], scientific = FALSE, digits = 4), "\n")
  cat("95% CI: [", round(finke_results$ci_lower[i], 3), ", ",
      round(finke_results$ci_upper[i], 3), "]\n")
  cat("Significant: ", ifelse(finke_results$significant[i], "YES (p < 0.05)", "NO"), "\n")
}
```

```
##
##  Humans vs o3
## ----------------------------------------
## Proportions:  0.625  vs  0.646
## Difference:  -0.021
## Chi-squared:  0.127
## Degrees of freedom:  1
## P-value:  0.7216
## 95% CI: [ -0.114 ,  0.073 ]
## Significant:  NO
##
##  Humans vs o3-GPT-Image
## ----------------------------------------
## Proportions:  0.625  vs  0.522
## Difference:  0.103
## Chi-squared:  2.202
## Degrees of freedom:  1
## P-value:  0.1378
## 95% CI: [ -0.034 ,  0.241 ]
## Significant:  NO
##
```

```
##  Humans vs o3-Pro
## ----------------------------------------
## Proportions:  0.625  vs  0.672
## Difference:  -0.047
## Chi-squared:  0.865
## Degrees of freedom:  1
## P-value:  0.3525
## 95% CI: [ -0.139 ,  0.045 ]
## Significant:  NO
##
##  Humans vs GPT-4.1
## ----------------------------------------
## Proportions:  0.625  vs  0.469
## Difference:  0.157
## Chi-squared:  10.863
## Degrees of freedom:  1
## P-value:  0.0009808
## 95% CI: [ 0.06 ,  0.254 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.625  vs  0.335
## Difference:  0.29
## Chi-squared:  37.921
## Degrees of freedom:  1
## P-value:  0.0000000007367
## 95% CI: [ 0.198 ,  0.382 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.625  vs  0.411
## Difference:  0.214
## Chi-squared:  20.523
## Degrees of freedom:  1
## P-value:  0.000005893
## 95% CI: [ 0.118 ,  0.31 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o4-mini
## ----------------------------------------
## Proportions:  0.625  vs  0.473
## Difference:  0.152
## Chi-squared:  19.398
## Degrees of freedom:  1
## P-value:  0.00001061
## 95% CI: [ 0.082 ,  0.222 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.625  vs  0.496
## Difference:  0.13
## Chi-squared:  7.359
## Degrees of freedom:  1
## P-value:  0.006675
## 95% CI: [ 0.032 ,  0.227 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.625  vs  0.359
## Difference:  0.267
## Chi-squared:  32.005
## Degrees of freedom:  1
## P-value:  0.00000001538
## 95% CI: [ 0.173 ,  0.36 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.625  vs  0.348
## Difference:  0.277
## Chi-squared:  17.555
## Degrees of freedom:  1
## P-value:  0.00002791
## 95% CI: [ 0.145 ,  0.409 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs All-OpenAI
## ----------------------------------------
## Proportions:  0.625  vs  0.499
## Difference:  0.126
## Chi-squared:  36.591
## Degrees of freedom:  1
## P-value:  0.000000001457
## 95% CI: [ 0.085 ,  0.168 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.625  vs  0.433
## Difference:  0.193
## Chi-squared:  64.624
## Degrees of freedom:  1
## P-value:  0.0000000000000009064
## 95% CI: [ 0.145 ,  0.241 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs All-Gemini
## ----------------------------------------
## Proportions:  0.625  vs  0.412
## Difference:  0.214
## Chi-squared:  46.53
## Degrees of freedom:  1
## P-value:  0.000000000009023
## 95% CI: [ 0.151 ,  0.277 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o3-GPT-Image
## ----------------------------------------
## Proportions:  0.646  vs  0.522
## Difference:  0.124
## Chi-squared:  2.089
## Degrees of freedom:  1
## P-value:  0.1484
## 95% CI: [ -0.041 ,  0.289 ]
## Significant:  NO
##
##  o3 vs o3-Pro
## ----------------------------------------
## Proportions:  0.646  vs  0.672
## Difference:  -0.026
## Chi-squared:  0.086
## Degrees of freedom:  1
## P-value:  0.7696
## 95% CI: [ -0.154 ,  0.102 ]
## Significant:  NO
##
##  o3 vs GPT-4.1
## ----------------------------------------
## Proportions:  0.646  vs  0.469
## Difference:  0.177
## Chi-squared:  6.957
## Degrees of freedom:  1
## P-value:  0.008349
## 95% CI: [ 0.045 ,  0.309 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.646  vs  0.335
## Difference:  0.311
## Chi-squared:  21.969
## Degrees of freedom:  1
## P-value:  0.000002771
## 95% CI: [ 0.182 ,  0.439 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.646  vs  0.411
## Difference:  0.235
## Chi-squared:  12.354
## Degrees of freedom:  1
## P-value:  0.0004399
## 95% CI: [ 0.104 ,  0.366 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o4-mini
## ----------------------------------------
## Proportions:  0.646  vs  0.473
## Difference:  0.173
## Chi-squared:  8.908
## Degrees of freedom:  1
## P-value:  0.002839
## 95% CI: [ 0.06 ,  0.285 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.646  vs  0.496
## Difference:  0.15
## Chi-squared:  4.945
## Degrees of freedom:  1
## P-value:  0.02617
## 95% CI: [ 0.018 ,  0.283 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.646  vs  0.359
## Difference:  0.287
## Chi-squared:  18.694
## Degrees of freedom:  1
## P-value:  0.00001535
## 95% CI: [ 0.158 ,  0.417 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.646  vs  0.348
## Difference:  0.298
## Chi-squared:  13.143
## Degrees of freedom:  1
## P-value:  0.0002886
## 95% CI: [ 0.138 ,  0.458 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs All-OpenAI
## ----------------------------------------
## Proportions:  0.646  vs  0.499
## Difference:  0.147
## Chi-squared:  8.616
## Degrees of freedom:  1
## P-value:  0.003333
## 95% CI: [ 0.051 ,  0.244 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.646  vs  0.433
## Difference:  0.214
## Chi-squared:  17.5
## Degrees of freedom:  1
## P-value:  0.00002874
## 95% CI: [ 0.114 ,  0.313 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs All-Gemini
## ----------------------------------------
## Proportions:  0.646  vs  0.412
## Difference:  0.235
## Chi-squared:  17.992
## Degrees of freedom:  1
## P-value:  0.00002218
## 95% CI: [ 0.127 ,  0.343 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs o3-Pro
## ----------------------------------------
## Proportions:  0.522  vs  0.672
## Difference:  -0.15
## Chi-squared:  3.24
## Degrees of freedom:  1
## P-value:  0.07186
## 95% CI: [ -0.315 ,  0.014 ]
## Significant:  NO
##
##  o3-GPT-Image vs GPT-4.1
## ----------------------------------------
## Proportions:  0.522  vs  0.469
## Difference:  0.053
## Chi-squared:  0.266
## Degrees of freedom:  1
## P-value:  0.6062
## 95% CI: [ -0.114 ,  0.22 ]
## Significant:  NO
##
##  o3-GPT-Image vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.522  vs  0.335
## Difference:  0.187
## Chi-squared:  5.062
## Degrees of freedom:  1
## P-value:  0.02445
## 95% CI: [ 0.022 ,  0.351 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.522  vs  0.411
## Difference:  0.111
## Chi-squared:  1.557
## Degrees of freedom:  1
## P-value:  0.2121
## 95% CI: [ -0.056 ,  0.277 ]
## Significant:  NO
##
```

```
##  o3-GPT-Image vs o4-mini
## ----------------------------------------
## Proportions:  0.522  vs  0.473
## Difference:  0.049
## Chi-squared:  0.28
## Degrees of freedom:  1
## P-value:  0.597
## 95% CI: [ -0.103 ,  0.2 ]
## Significant:  NO
##
##  o3-GPT-Image vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.522  vs  0.496
## Difference:  0.026
## Chi-squared:  0.03
## Degrees of freedom:  1
## P-value:  0.8625
## 95% CI: [ -0.141 ,  0.194 ]
## Significant:  NO
##
##  o3-GPT-Image vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.522  vs  0.359
## Difference:  0.163
## Chi-squared:  3.746
## Degrees of freedom:  1
## P-value:  0.05293
## 95% CI: [ -0.002 ,  0.328 ]
## Significant:  NO
##
##  o3-GPT-Image vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.522  vs  0.348
## Difference:  0.174
## Chi-squared:  3.006
## Degrees of freedom:  1
## P-value:  0.08294
## 95% CI: [ -0.018 ,  0.365 ]
## Significant:  NO
##
##  o3-GPT-Image vs All-OpenAI
## ----------------------------------------
## Proportions:  0.522  vs  0.499
## Difference:  0.023
## Chi-squared:  0.045
## Degrees of freedom:  1
## P-value:  0.8318
## 95% CI: [ -0.116 ,  0.162 ]
## Significant:  NO
##
##  o3-GPT-Image vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.522  vs  0.433
## Difference:  0.09
## Chi-squared:  1.429
## Degrees of freedom:  1
## P-value:  0.232
## 95% CI: [ -0.052 ,  0.231 ]
## Significant:  NO
##
##  o3-GPT-Image vs All-Gemini
## ----------------------------------------
## Proportions:  0.522  vs  0.412
## Difference:  0.11
## Chi-squared:  2.06
## Degrees of freedom:  1
## P-value:  0.1513
## 95% CI: [ -0.038 ,  0.259 ]
## Significant:  NO
##
##  o3-Pro vs GPT-4.1
## ----------------------------------------
## Proportions:  0.672  vs  0.469
## Difference:  0.204
## Chi-squared:  9.348
## Degrees of freedom:  1
## P-value:  0.002232
## 95% CI: [ 0.073 ,  0.335 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.672  vs  0.335
## Difference:  0.337
## Chi-squared:  25.94
## Degrees of freedom:  1
## P-value:  0.0000003522
## 95% CI: [ 0.21 ,  0.465 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.672  vs  0.411
## Difference:  0.261
## Chi-squared:  15.441
## Degrees of freedom:  1
## P-value:  0.0000851
## 95% CI: [ 0.131 ,  0.391 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o4-mini
## ----------------------------------------
## Proportions:  0.672  vs  0.473
## Difference:  0.199
## Chi-squared:  11.968
## Degrees of freedom:  1
## P-value:  0.0005412
## 95% CI: [ 0.088 ,  0.31 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.672  vs  0.496
## Difference:  0.177
## Chi-squared:  6.999
## Degrees of freedom:  1
## P-value:  0.008156
## 95% CI: [ 0.046 ,  0.308 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.672  vs  0.359
## Difference:  0.314
## Chi-squared:  22.397
## Degrees of freedom:  1
## P-value:  0.000002217
## 95% CI: [ 0.185 ,  0.442 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.672  vs  0.348
## Difference:  0.324
## Chi-squared:  15.796
## Degrees of freedom:  1
## P-value:  0.00007053
## 95% CI: [ 0.165 ,  0.484 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs All-OpenAI
## ----------------------------------------
## Proportions:  0.672  vs  0.499
## Difference:  0.174
## Chi-squared:  12.086
## Degrees of freedom:  1
## P-value:  0.000508
## 95% CI: [ 0.079 ,  0.268 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.672  vs  0.433
## Difference:  0.24
## Chi-squared:  22.155
## Degrees of freedom:  1
## P-value:  0.000002515
## 95% CI: [ 0.142 ,  0.338 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs All-Gemini
## ----------------------------------------
## Proportions:  0.672  vs  0.412
## Difference:  0.261
## Chi-squared:  22.332
## Degrees of freedom:  1
## P-value:  0.000002293
## 95% CI: [ 0.154 ,  0.368 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.469  vs  0.335
## Difference:  0.133
## Chi-squared:  3.902
## Degrees of freedom:  1
## P-value:  0.04823
## 95% CI: [ 0.002 ,  0.265 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.469  vs  0.411
## Difference:  0.057
## Chi-squared:  0.586
## Degrees of freedom:  1
## P-value:  0.4441
## 95% CI: [ -0.076 ,  0.191 ]
## Significant:  NO
##
##  GPT-4.1 vs o4-mini
## ----------------------------------------
## Proportions:  0.469  vs  0.473
## Difference:  -0.005
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.119 ,  0.109 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.469  vs  0.496
## Difference:  -0.027
## Chi-squared:  0.084
## Degrees of freedom:  1
## P-value:  0.7718
## 95% CI: [ -0.162 ,  0.108 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.469  vs  0.359
## Difference:  0.11
## Chi-squared:  2.555
## Degrees of freedom:  1
## P-value:  0.11
## 95% CI: [ -0.022 ,  0.242 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.469  vs  0.348
## Difference:  0.12
## Chi-squared:  1.901
## Degrees of freedom:  1
## P-value:  0.168
## 95% CI: [ -0.042 ,  0.283 ]
## Significant:  NO
##
##  GPT-4.1 vs All-OpenAI
## ----------------------------------------
## Proportions:  0.469  vs  0.499
## Difference:  -0.03
## Chi-squared:  0.275
## Degrees of freedom:  1
## P-value:  0.6002
## 95% CI: [ -0.13 ,  0.07 ]
## Significant:  NO
##
##  GPT-4.1 vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.469  vs  0.433
## Difference:  0.036
## Chi-squared:  0.397
## Degrees of freedom:  1
## P-value:  0.5285
## 95% CI: [ -0.066 ,  0.139 ]
## Significant:  NO
##
##  GPT-4.1 vs All-Gemini
## ----------------------------------------
## Proportions:  0.469  vs  0.412
## Difference:  0.057
## Chi-squared:  0.925
## Degrees of freedom:  1
## P-value:  0.3361
## 95% CI: [ -0.054 ,  0.168 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.335  vs  0.411
## Difference:  -0.076
## Chi-squared:  1.174
## Degrees of freedom:  1
## P-value:  0.2786
## 95% CI: [ -0.206 ,  0.054 ]
## Significant:  NO
##
```

```
##  GPT-4.1-GPT-Image vs o4-mini
## ----------------------------------------
## Proportions:  0.335  vs  0.473
## Difference:  -0.138
## Chi-squared:  5.679
## Degrees of freedom:  1
## P-value:  0.01717
## 95% CI: [ -0.25 ,  -0.026 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.335  vs  0.496
## Difference:  -0.16
## Chi-squared:  5.714
## Degrees of freedom:  1
## P-value:  0.01683
## 95% CI: [ -0.292 ,  -0.029 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.335  vs  0.359
## Difference:  -0.023
## Chi-squared:  0.06
## Degrees of freedom:  1
## P-value:  0.8062
## 95% CI: [ -0.152 ,  0.105 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.335  vs  0.348
## Difference:  -0.013
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  0.9947
## 95% CI: [ -0.173 ,  0.147 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs All-OpenAI
## ----------------------------------------
## Proportions:  0.335  vs  0.499
## Difference:  -0.164
## Chi-squared:  10.703
## Degrees of freedom:  1
## P-value:  0.00107
## 95% CI: [ -0.259 ,  -0.068 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.335  vs  0.433
## Difference:  -0.097
## Chi-squared:  3.49
## Degrees of freedom:  1
## P-value:  0.06173
## 95% CI: [ -0.195 ,  0.001 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs All-Gemini
## ----------------------------------------
## Proportions:  0.335  vs  0.412
## Difference:  -0.076
## Chi-squared:  1.781
## Degrees of freedom:  1
## P-value:  0.182
## 95% CI: [ -0.183 ,  0.031 ]
## Significant:  NO
##
##  ChatGPT-4o vs o4-mini
## ----------------------------------------
## Proportions:  0.411  vs  0.473
## Difference:  -0.062
## Chi-squared:  1.006
## Degrees of freedom:  1
## P-value:  0.3159
## 95% CI: [ -0.177 ,  0.053 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.411  vs  0.496
## Difference:  -0.084
## Chi-squared:  1.402
## Degrees of freedom:  1
## P-value:  0.2365
## 95% CI: [ -0.218 ,  0.049 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.411  vs  0.359
## Difference:  0.053
## Chi-squared:  0.496
## Degrees of freedom:  1
## P-value:  0.4812
## 95% CI: [ -0.079 ,  0.184 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.411  vs  0.348
## Difference:  0.063
## Chi-squared:  0.429
## Degrees of freedom:  1
## P-value:  0.5127
## 95% CI: [ -0.099 ,  0.225 ]
## Significant:  NO
##
##  ChatGPT-4o vs All-OpenAI
## ---------------------------------------
## Proportions:  0.411  vs  0.499
## Difference:  -0.088
## Chi-squared:  2.908
## Degrees of freedom:  1
## P-value:  0.08812
## 95% CI: [ -0.186 ,  0.011 ]
## Significant:  NO
##
##  ChatGPT-4o vs Other-OpenAI
## ---------------------------------------
## Proportions:  0.411  vs  0.433
## Difference:  -0.021
## Chi-squared:  0.106
## Degrees of freedom:  1
## P-value:  0.7449
## 95% CI: [ -0.123 ,  0.08 ]
## Significant:  NO
##
##  ChatGPT-4o vs All-Gemini
## ---------------------------------------
## Proportions:  0.411  vs  0.412
## Difference:  0
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.104 ,  0.104 ]
## Significant:  NO
##
##  o4-mini vs Gemini-2.5
## ---------------------------------------
## Proportions:  0.473  vs  0.496
## Difference:  -0.022
## Chi-squared:  0.083
## Degrees of freedom:  1
## P-value:  0.773
## 95% CI: [ -0.138 ,  0.093 ]
## Significant:  NO
##
##  o4-mini vs Gemini-2.0-Flash
## ---------------------------------------
## Proportions:  0.473  vs  0.359
## Difference:  0.115
## Chi-squared:  3.824
## Degrees of freedom:  1
## P-value:  0.05053
## 95% CI: [ 0.002 ,  0.227 ]
## Significant:  NO
##
##  o4-mini vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.473  vs  0.348
## Difference:  0.125
## Chi-squared:  2.551
## Degrees of freedom:  1
## P-value:  0.1103
## 95% CI: [ -0.021 ,  0.272 ]
## Significant:  NO
##
##  o4-mini vs All-OpenAI
## ----------------------------------------
## Proportions:  0.473  vs  0.499
## Difference:  -0.026
## Chi-squared:  0.396
## Degrees of freedom:  1
## P-value:  0.529
## 95% CI: [ -0.099 ,  0.048 ]
## Significant:  NO
##
##  o4-mini vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.473  vs  0.433
## Difference:  0.041
## Chi-squared:  1.005
## Degrees of freedom:  1
## P-value:  0.3161
## 95% CI: [ -0.037 ,  0.118 ]
## Significant:  NO
##
##  o4-mini vs All-Gemini
## ----------------------------------------
## Proportions:  0.473  vs  0.412
## Difference:  0.062
## Chi-squared:  1.832
## Degrees of freedom:  1
## P-value:  0.1759
## 95% CI: [ -0.026 ,  0.15 ]
## Significant:  NO
##
##  Gemini-2.5 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.496  vs  0.359
## Difference:  0.137
## Chi-squared:  4.059
## Degrees of freedom:  1
## P-value:  0.04393
## 95% CI: [ 0.005 ,  0.269 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5 vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.496  vs  0.348
## Difference:  0.147
## Chi-squared:  2.946
## Degrees of freedom:  1
## P-value:  0.08609
## 95% CI: [ -0.015 ,  0.31 ]
## Significant:  NO
##
##  Gemini-2.5 vs All-OpenAI
## ----------------------------------------
## Proportions:  0.496  vs  0.499
## Difference:  -0.003
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.102 ,  0.095 ]
## Significant:  NO
##
##  Gemini-2.5 vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.496  vs  0.433
## Difference:  0.063
## Chi-squared:  1.378
## Degrees of freedom:  1
## P-value:  0.2404
## 95% CI: [ -0.04 ,  0.166 ]
## Significant:  NO
##
##  Gemini-2.5 vs All-Gemini
## ----------------------------------------
## Proportions:  0.496  vs  0.412
## Difference:  0.084
## Chi-squared:  2.146
## Degrees of freedom:  1
## P-value:  0.143
## 95% CI: [ -0.027 ,  0.195 ]
## Significant:  NO
##
##  Gemini-2.0-Flash vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.359  vs  0.348
## Difference:  0.01
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.148 ,  0.169 ]
## Significant:  NO
##
##  Gemini-2.0-Flash vs All-OpenAI
## ----------------------------------------
## Proportions:  0.359  vs  0.499
## Difference:  -0.14
## Chi-squared:  7.777
## Degrees of freedom:  1
## P-value:  0.005292
## 95% CI: [ -0.237 ,  -0.044 ]
## Significant:  YES (p < 0.05)
##
```

```
##  Gemini-2.0-Flash vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.359  vs  0.433
## Difference:  -0.074
## Chi-squared:  1.936
## Degrees of freedom:  1
## P-value:  0.1641
## 95% CI: [ -0.173 ,  0.026 ]
## Significant:  NO
##
##  Gemini-2.0-Flash vs All-Gemini
## ----------------------------------------
## Proportions:  0.359  vs  0.412
## Difference:  -0.053
## Chi-squared:  0.787
## Degrees of freedom:  1
## P-value:  0.3749
## 95% CI: [ -0.161 ,  0.055 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-GPT-Image vs All-OpenAI
## ----------------------------------------
## Proportions:  0.348  vs  0.499
## Difference:  -0.151
## Chi-squared:  4.518
## Degrees of freedom:  1
## P-value:  0.03354
## 95% CI: [ -0.284 ,  -0.017 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-GPT-Image vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.348  vs  0.433
## Difference:  -0.084
## Chi-squared:  1.253
## Degrees of freedom:  1
## P-value:  0.2629
## 95% CI: [ -0.22 ,  0.052 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-GPT-Image vs All-Gemini
## ----------------------------------------
## Proportions:  0.348  vs  0.412
## Difference:  -0.063
## Chi-squared:  0.588
## Degrees of freedom:  1
## P-value:  0.4433
## 95% CI: [ -0.206 ,  0.08 ]
## Significant:  NO
##
##  All-OpenAI vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.499  vs  0.433
## Difference:  0.066
## Chi-squared:  6.117
## Degrees of freedom:  1
## P-value:  0.01339
## 95% CI: [ 0.014 ,  0.119 ]
## Significant:  YES (p < 0.05)
##
##  All-OpenAI vs All-Gemini
## ----------------------------------------
## Proportions:  0.499  vs  0.412
## Difference:  0.087
## Chi-squared:  6.548
## Degrees of freedom:  1
## P-value:  0.0105
## 95% CI: [ 0.021 ,  0.154 ]
## Significant:  YES (p < 0.05)
##
##  Other-OpenAI vs All-Gemini
## ----------------------------------------
## Proportions:  0.433  vs  0.412
## Difference:  0.021
## Chi-squared:  0.279
## Degrees of freedom:  1
## P-value:  0.5974
## 95% CI: [ -0.05 ,  0.092 ]
## Significant:  NO
```

```r
# Summary table
finke_summary <- finke_results %>%
  select(comparison, diff, chi_squared, p_value, significant) %>%
  mutate(diff = round(diff, 3),
         p_value = round(p_value, 4))

cat("\n\nSummary Table - Finke Tasks:\n")
```

```
##
##
## Summary Table - Finke Tasks:
```

```r
print(kable(finke_summary, format = "simple"))
```

```
##
##
##               comparison                                            diff   chi_squared   p_value   sig
## ------------  ------------------------------------------------  -------  ------------  --------  ---
## X-squared     Humans vs o3                                       -0.021     0.1269843    0.7216  FALS
## X-squared1    Humans vs o3-GPT-Image                              0.103     2.2022705    0.1378  FALS
## X-squared2    Humans vs o3-Pro                                   -0.047     0.8645532    0.3525  FALS
## X-squared3    Humans vs GPT-4.1                                   0.157    10.8634029    0.0010  TRU
## X-squared4    Humans vs GPT-4.1-GPT-Image                         0.290    37.9209213    0.0000  TRU
## X-squared5    Humans vs ChatGPT-4o                                0.214    20.5226545    0.0000  TRU
## X-squared6    Humans vs o4-mini                                   0.152    19.3982559    0.0000  TRU
## X-squared7    Humans vs Gemini-2.5                                0.130     7.3585182    0.0067  TRU
## X-squared8    Humans vs Gemini-2.0-Flash                          0.267    32.0046091    0.0000  TRU
## X-squared9    Humans vs Gemini-2.0-Flash-GPT-Image                0.277    17.5551447    0.0000  TRU
## X-squared10   Humans vs All-OpenAI                                0.126    36.5910627    0.0000  TRU
## X-squared11   Humans vs Other-OpenAI                              0.193    64.6241507    0.0000  TRU
## X-squared12   Humans vs All-Gemini                                0.214    46.5298252    0.0000  TRU
## X-squared13   o3 vs o3-GPT-Image                                  0.124     2.0890892    0.1484  FALS
## X-squared14   o3 vs o3-Pro                                       -0.026     0.0857545    0.7696  FALS
## X-squared15   o3 vs GPT-4.1                                       0.177     6.9570149    0.0083  TRU
## X-squared16   o3 vs GPT-4.1-GPT-Image                             0.311    21.9688840    0.0000  TRU
## X-squared17   o3 vs ChatGPT-4o                                    0.235    12.3544757    0.0004  TRU
## X-squared18   o3 vs o4-mini                                       0.173     8.9084677    0.0028  TRU
## X-squared19   o3 vs Gemini-2.5                                    0.150     4.9448670    0.0262  TRU
## X-squared20   o3 vs Gemini-2.0-Flash                              0.287    18.6940156    0.0000  TRU
## X-squared21   o3 vs Gemini-2.0-Flash-GPT-Image                    0.298    13.1427658    0.0003  TRU
## X-squared22   o3 vs All-OpenAI                                    0.147     8.6156814    0.0033  TRU
## X-squared23   o3 vs Other-OpenAI                                  0.214    17.4997126    0.0000  TRU
## X-squared24   o3 vs All-Gemini                                    0.235    17.9921560    0.0000  TRU
## X-squared25   o3-GPT-Image vs o3-Pro                             -0.150     3.2399638    0.0719  FALS
## X-squared26   o3-GPT-Image vs GPT-4.1                             0.053     0.2656757    0.6062  FALS
## X-squared27   o3-GPT-Image vs GPT-4.1-GPT-Image                   0.187     5.0623802    0.0245  TRU
## X-squared28   o3-GPT-Image vs ChatGPT-4o                          0.111     1.5569094    0.2121  FALS
## X-squared29   o3-GPT-Image vs o4-mini                             0.049     0.2795365    0.5970  FALS
## X-squared30   o3-GPT-Image vs Gemini-2.5                          0.026     0.0299912    0.8625  FALS
## X-squared31   o3-GPT-Image vs Gemini-2.0-Flash                    0.163     3.7462392    0.0529  FALS
## X-squared32   o3-GPT-Image vs Gemini-2.0-Flash-GPT-Image          0.174     3.0062521    0.0829  FALS
## X-squared33   o3-GPT-Image vs All-OpenAI                          0.023     0.0451057    0.8318  FALS
## X-squared34   o3-GPT-Image vs Other-OpenAI                        0.090     1.4285701    0.2320  FALS
## X-squared35   o3-GPT-Image vs All-Gemini                          0.110     2.0595050    0.1513  FALS
## X-squared36   o3-Pro vs GPT-4.1                                   0.204     9.3479304    0.0022  TRU
## X-squared37   o3-Pro vs GPT-4.1-GPT-Image                         0.337    25.9400008    0.0000  TRU
## X-squared38   o3-Pro vs ChatGPT-4o                                0.261    15.4414964    0.0001  TRU
## X-squared39   o3-Pro vs o4-mini                                   0.199    11.9681354    0.0005  TRU
## X-squared40   o3-Pro vs Gemini-2.5                                0.177     6.9989908    0.0082  TRU
## X-squared41   o3-Pro vs Gemini-2.0-Flash                          0.314    22.3973920    0.0000  TRU
## X-squared42   o3-Pro vs Gemini-2.0-Flash-GPT-Image                0.324    15.7964958    0.0001  TRU
## X-squared43   o3-Pro vs All-OpenAI                                0.174    12.0862146    0.0005  TRU
## X-squared44   o3-Pro vs Other-OpenAI                              0.240    22.1547317    0.0000  TRU
## X-squared45   o3-Pro vs All-Gemini                                0.261    22.3323527    0.0000  TRU
## X-squared46   GPT-4.1 vs GPT-4.1-GPT-Image                        0.133     3.9019637    0.0482  TRU
## X-squared47   GPT-4.1 vs ChatGPT-4o                               0.057     0.5857677    0.4441  FALS
## X-squared48   GPT-4.1 vs o4-mini                                 -0.005     0.0000000    1.0000  FALS
## X-squared49   GPT-4.1 vs Gemini-2.5                              -0.027     0.0841235    0.7718  FALS
## X-squared50   GPT-4.1 vs Gemini-2.0-Flash                         0.110     2.5546422    0.1100  FALS
## X-squared51   GPT-4.1 vs Gemini-2.0-Flash-GPT-Image               0.120     1.9006799    0.1680  FALS
## X-squared52   GPT-4.1 vs All-OpenAI                              -0.030     0.2746521    0.6002  FALS
## X-squared53   GPT-4.1 vs Other-OpenAI                             0.036     0.3973194    0.5285  FALS
## X-squared54   GPT-4.1 vs All-Gemini                               0.057     0.9253629    0.3361  FALS
## X-squared55   GPT-4.1-GPT-Image vs ChatGPT-4o                    -0.076     1.1738923    0.2786  FALS
## X-squared56   GPT-4.1-GPT-Image vs o4-mini                       -0.138     5.6788151    0.0172  TRU
## X-squared57   GPT-4.1-GPT-Image vs Gemini-2.5                    -0.160     5.7137496    0.0168  TRU
## X-squared58   GPT-4.1-GPT-Image vs Gemini-2.0-Flash              -0.023     0.0602008    0.8062  FALS
## X-squared59   GPT-4.1-GPT-Image vs Gemini-2.0-Flash-GPT-Image    -0.013     0.0000439    0.9947  FALS
## X-squared60   GPT-4.1-GPT-Image vs All-OpenAI                    -0.164    10.7029291    0.0011  TRU
## X-squared61   GPT-4.1-GPT-Image vs Other-OpenAI                  -0.097     3.4903539    0.0617  FALS
## X-squared62   GPT-4.1-GPT-Image vs All-Gemini                    -0.076     1.7808283    0.1820  FALS
## X-squared63   ChatGPT-4o vs o4-mini                              -0.062     1.0058158    0.3159  FALS
## X-squared64   ChatGPT-4o vs Gemini-2.5                           -0.084     1.4016312    0.2365  FALS
## X-squared65   ChatGPT-4o vs Gemini-2.0-Flash                    0.053     0.4960368    0.4812  FALS
## X-squared66   ChatGPT-4o vs Gemini-2.0-Flash-GPT-Image          0.063     0.4285640    0.5127  FALS
## X-squared67   ChatGPT-4o vs All-OpenAI                         -0.088     2.9083158    0.0881  FALS
## X-squared68   ChatGPT-4o vs Other-OpenAI                       -0.021     0.1058598    0.7449  FALS
## X-squared69   ChatGPT-4o vs All-Gemini                          0.000     0.0000000    1.0000  FALS
## X-squared70   o4-mini vs Gemini-2.5                            -0.022     0.0831828    0.7730  FALS
## X-squared71   o4-mini vs Gemini-2.0-Flash                       0.115     3.8237939    0.0505  FALS
## X-squared72   o4-mini vs Gemini-2.0-Flash-GPT-Image             0.125     2.5505536    0.1103  FALS
## X-squared73   o4-mini vs All-OpenAI                            -0.026     0.3963011    0.5290  FALS
## X-squared74   o4-mini vs Other-OpenAI                           0.041     1.0049572    0.3161  FALS
## X-squared75   o4-mini vs All-Gemini                             0.062     1.8321593    0.1759  FALS
## X-squared76   Gemini-2.5 vs Gemini-2.0-Flash                    0.137     4.0592143    0.0439  TRU
## X-squared77   Gemini-2.5 vs Gemini-2.0-Flash-GPT-Image          0.147     2.9460588    0.0861  FALS
## X-squared78   Gemini-2.5 vs All-OpenAI                         -0.003     0.0000000    1.0000  FALS
## X-squared79   Gemini-2.5 vs Other-OpenAI                        0.063     1.3784285    0.2404  FALS
## X-squared80   Gemini-2.5 vs All-Gemini                          0.084     2.1457738    0.1430  FALS
## X-squared81   Gemini-2.0-Flash vs Gemini-2.0-Flash-GPT-Image    0.010     0.0000000    1.0000  FALS
## X-squared82   Gemini-2.0-Flash vs All-OpenAI                   -0.140     7.7769940    0.0053  TRU
## X-squared83   Gemini-2.0-Flash vs Other-OpenAI                 -0.074     1.9361967    0.1641  FALS
## X-squared84   Gemini-2.0-Flash vs All-Gemini                   -0.053     0.7874168    0.3749  FALS
## X-squared85   Gemini-2.0-Flash-GPT-Image vs All-OpenAI         -0.151     4.5180985    0.0335  TRU
## X-squared86   Gemini-2.0-Flash-GPT-Image vs Other-OpenAI       -0.084     1.2533672    0.2629  FALS
## X-squared87   Gemini-2.0-Flash-GPT-Image vs All-Gemini         -0.063     0.5876921    0.4433  FALS
## X-squared88   All-OpenAI vs Other-OpenAI                        0.066     6.1167734    0.0134  TRU
## X-squared89   All-OpenAI vs All-Gemini                          0.087     6.5478219    0.0105  TRU
## X-squared90   Other-OpenAI vs All-Gemini                        0.021     0.2789325    0.5974  FALS
```

48 Novel Tasks - All Pairwise Comparisons

```r
# Test all combinations for 48 Novel tasks
novel_48_results <- test_all_combinations(novel_data, "48 Novel")

# Display results
cat("All Pairwise Comparisons for 48 Novel Tasks:\n")
```

```
## All Pairwise Comparisons for 48 Novel Tasks:
```

```r
cat(paste(rep("=", 80), collapse = ""), "\n")
```

```
## ================================================================================
```

```r
for (i in 1:nrow(novel_48_results)) {
  cat("\n", novel_48_results$comparison[i], "\n")
  cat(paste(rep("-", 40), collapse = ""), "\n")
  cat("Proportions: ", round(novel_48_results$prop1[i], 3), " vs ",
      round(novel_48_results$prop2[i], 3), "\n")
  cat("Difference: ", round(novel_48_results$diff[i], 3), "\n")
  cat("Chi-squared: ", round(novel_48_results$chi_squared[i], 3), "\n")
  cat("Degrees of freedom: ", round(novel_48_results$df[i], 3), "\n")
  cat("P-value: ", format(novel_48_results$p_value[i], scientific = FALSE, digits = 4), "\n")
  cat("95% CI: [", round(novel_48_results$ci_lower[i], 3), ", ",
      round(novel_48_results$ci_upper[i], 3), "]\n")
  cat("Significant: ", ifelse(novel_48_results$significant[i], "YES (p < 0.05)", "NO"), "\n")
}
```

```
##
##  Humans vs o3
## ----------------------------------------
## Proportions:  0.517  vs  0.591
## Difference:  -0.073
## Chi-squared:  9.265
## Degrees of freedom:  1
## P-value:  0.002336
## 95% CI: [ -0.12 ,  -0.026 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o3-GPT-Image
## ----------------------------------------
## Proportions:  0.517  vs  0.546
## Difference:  -0.028
## Chi-squared:  0.629
## Degrees of freedom:  1
## P-value:  0.4277
## 95% CI: [ -0.095 ,  0.038 ]
## Significant:  NO
##
##  Humans vs o3-Pro
## ----------------------------------------
## Proportions:  0.517  vs  0.628
## Difference:  -0.11
## Chi-squared:  21.247
## Degrees of freedom:  1
## P-value:  0.000004037
## 95% CI: [ -0.157 ,  -0.064 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-4.1
## ----------------------------------------
## Proportions:  0.517  vs  0.404
## Difference:  0.113
## Chi-squared:  22.409
## Degrees of freedom:  1
## P-value:  0.000002203
## 95% CI: [ 0.067 ,  0.16 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.517  vs  0.397
## Difference:  0.12
## Chi-squared:  25.169
## Degrees of freedom:  1
## P-value:  0.0000005251
## 95% CI: [ 0.073 ,  0.167 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.517  vs  0.417
## Difference:  0.1
## Chi-squared:  17.303
## Degrees of freedom:  1
## P-value:  0.00003186
## 95% CI: [ 0.053 ,  0.147 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o4-mini
## ----------------------------------------
## Proportions:  0.517  vs  0.501
## Difference:  0.016
## Chi-squared:  0.773
## Degrees of freedom:  1
## P-value:  0.3794
## 95% CI: [ -0.019 ,  0.051 ]
## Significant:  NO
##
##  Humans vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.517  vs  0.445
## Difference:  0.072
## Chi-squared:  8.905
## Degrees of freedom:  1
## P-value:  0.002843
## 95% CI: [ 0.025 ,  0.119 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.517  vs  0.394
## Difference:  0.123
## Chi-squared:  26.487
## Degrees of freedom:  1
## P-value:  0.0000002653
## 95% CI: [ 0.077 ,  0.17 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.517  vs  0.302
## Difference:  0.215
## Chi-squared:  41.98
## Degrees of freedom:  1
## P-value:  0.00000000009221
## 95% CI: [ 0.154 ,  0.277 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs All-OpenAI
## ----------------------------------------
## Proportions:  0.517  vs  0.495
## Difference:  0.022
## Chi-squared:  4.384
## Degrees of freedom:  1
## P-value:  0.03627
## 95% CI: [ 0.001 ,  0.043 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.517  vs  0.444
## Difference:  0.073
## Chi-squared:  36.206
## Degrees of freedom:  1
## P-value:  0.000000001775
## 95% CI: [ 0.049 ,  0.097 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs All-Gemini
## ----------------------------------------
## Proportions:  0.517  vs  0.396
## Difference:  0.121
## Chi-squared:  58.142
## Degrees of freedom:  1
## P-value:  0.00000000000002439
## 95% CI: [ 0.09 ,  0.152 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o3-GPT-Image
## ----------------------------------------
## Proportions:  0.591  vs  0.546
## Difference:  0.045
## Chi-squared:  1.148
## Degrees of freedom:  1
## P-value:  0.2839
## 95% CI: [ -0.035 ,  0.125 ]
## Significant:  NO
##
##  o3 vs o3-Pro
## ----------------------------------------
## Proportions:  0.591  vs  0.628
## Difference:  -0.037
## Chi-squared:  1.234
## Degrees of freedom:  1
## P-value:  0.2666
## 95% CI: [ -0.101 ,  0.027 ]
## Significant:  NO
##
##  o3 vs GPT-4.1
## ----------------------------------------
## Proportions:  0.591  vs  0.404
## Difference:  0.187
## Chi-squared:  32.707
## Degrees of freedom:  1
## P-value:  0.00000001071
## 95% CI: [ 0.122 ,  0.251 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.591  vs  0.397
## Difference:  0.193
## Chi-squared:  35.135
## Degrees of freedom:  1
## P-value:  0.000000003076
## 95% CI: [ 0.129 ,  0.258 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.591  vs  0.417
## Difference:  0.173
## Chi-squared:  28.059
## Degrees of freedom:  1
## P-value:  0.0000001177
## 95% CI: [ 0.109 ,  0.237 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o4-mini
## ----------------------------------------
## Proportions:  0.591  vs  0.501
## Difference:  0.089
## Chi-squared:  9.855
## Degrees of freedom:  1
## P-value:  0.001693
## 95% CI: [ 0.033 ,  0.145 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.591  vs  0.445
## Difference:  0.145
## Chi-squared:  19.676
## Degrees of freedom:  1
## P-value:  0.000009176
## 95% CI: [ 0.081 ,  0.21 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.591  vs  0.394
## Difference:  0.196
## Chi-squared:  36.277
## Degrees of freedom:  1
## P-value:  0.000000001712
## 95% CI: [ 0.132 ,  0.261 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.591  vs  0.302
## Difference:  0.289
## Chi-squared:  52.19
## Degrees of freedom:  1
## P-value:  0.0000000000005038
## 95% CI: [ 0.213 ,  0.365 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs All-OpenAI
## ----------------------------------------
## Proportions:  0.591  vs  0.495
## Difference:  0.096
## Chi-squared:  15.096
## Degrees of freedom:  1
## P-value:  0.0001022
## 95% CI: [ 0.047 ,  0.144 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.591  vs  0.444
## Difference:  0.146
## Chi-squared:  33.783
## Degrees of freedom:  1
## P-value:  0.000000006161
## 95% CI: [ 0.097 ,  0.196 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs All-Gemini
## ----------------------------------------
## Proportions:  0.591  vs  0.396
## Difference:  0.194
## Chi-squared:  51.523
## Degrees of freedom:  1
## P-value:  0.0000000000007077
## 95% CI: [ 0.141 ,  0.248 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs o3-Pro
## ----------------------------------------
## Proportions:  0.546  vs  0.628
## Difference:  -0.082
## Chi-squared:  4.156
## Degrees of freedom:  1
## P-value:  0.04149
## 95% CI: [ -0.162 ,  -0.003 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs GPT-4.1
## ----------------------------------------
## Proportions:  0.546  vs  0.404
## Difference:  0.142
## Chi-squared:  12.402
## Degrees of freedom:  1
## P-value:  0.0004289
## 95% CI: [ 0.062 ,  0.222 ]
## Significant:  YES (p < 0.05)
```

```
##
##  o3-GPT-Image vs GPT-4.1-GPT-Image
## ---------------------------------------
## Proportions:  0.546  vs  0.397
## Difference:  0.148
## Chi-squared:  13.659
## Degrees of freedom:  1
## P-value:  0.0002192
## 95% CI: [ 0.069 ,  0.228 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs ChatGPT-4o
## ---------------------------------------
## Proportions:  0.546  vs  0.417
## Difference:  0.128
## Chi-squared:  10.05
## Degrees of freedom:  1
## P-value:  0.001523
## 95% CI: [ 0.048 ,  0.208 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs o4-mini
## ---------------------------------------
## Proportions:  0.546  vs  0.501
## Difference:  0.044
## Chi-squared:  1.325
## Degrees of freedom:  1
## P-value:  0.2496
## 95% CI: [ -0.029 ,  0.117 ]
## Significant:  NO
##
##  o3-GPT-Image vs Gemini-2.5
## ---------------------------------------
## Proportions:  0.546  vs  0.445
## Difference:  0.1
## Chi-squared:  6.036
## Degrees of freedom:  1
## P-value:  0.01402
## 95% CI: [ 0.02 ,  0.18 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Gemini-2.0-Flash
## ---------------------------------------
## Proportions:  0.546  vs  0.394
## Difference:  0.151
## Chi-squared:  14.256
## Degrees of freedom:  1
## P-value:  0.0001595
## 95% CI: [ 0.072 ,  0.231 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.546  vs  0.302
## Difference:  0.244
## Chi-squared:  28.191
## Degrees of freedom:  1
## P-value:  0.0000001099
## 95% CI: [ 0.154 ,  0.334 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs All-OpenAI
## ----------------------------------------
## Proportions:  0.546  vs  0.495
## Difference:  0.051
## Chi-squared:  2.104
## Degrees of freedom:  1
## P-value:  0.1469
## 95% CI: [ -0.017 ,  0.118 ]
## Significant:  NO
##
##  o3-GPT-Image vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.546  vs  0.444
## Difference:  0.101
## Chi-squared:  8.625
## Degrees of freedom:  1
## P-value:  0.003316
## 95% CI: [ 0.033 ,  0.17 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs All-Gemini
## ----------------------------------------
## Proportions:  0.546  vs  0.396
## Difference:  0.149
## Chi-squared:  17.7
## Degrees of freedom:  1
## P-value:  0.00002586
## 95% CI: [ 0.078 ,  0.221 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-4.1
## ----------------------------------------
## Proportions:  0.628  vs  0.404
## Difference:  0.224
## Chi-squared:  47.211
## Degrees of freedom:  1
## P-value:  0.000000000006374
## 95% CI: [ 0.16 ,  0.287 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.628  vs  0.397
## Difference:  0.23
## Chi-squared:  50.096
## Degrees of freedom:  1
## P-value:  0.000000000001464
## 95% CI: [ 0.167 ,  0.294 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.628  vs  0.417
## Difference:  0.21
## Chi-squared:  41.63
## Degrees of freedom:  1
## P-value:  0.0000000001103
## 95% CI: [ 0.146 ,  0.274 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o4-mini
## ----------------------------------------
## Proportions:  0.628  vs  0.501
## Difference:  0.126
## Chi-squared:  20.039
## Degrees of freedom:  1
## P-value:  0.000007586
## 95% CI: [ 0.071 ,  0.181 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.628  vs  0.445
## Difference:  0.182
## Chi-squared:  31.32
## Degrees of freedom:  1
## P-value:  0.00000002188
## 95% CI: [ 0.118 ,  0.246 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.628  vs  0.394
## Difference:  0.234
## Chi-squared:  51.448
## Degrees of freedom:  1
## P-value:  0.0000000000007354
## 95% CI: [ 0.17 ,  0.297 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.628  vs  0.302
## Difference:  0.326
## Chi-squared:  66.709
## Degrees of freedom:  1
## P-value:  0.0000000000000003146
## 95% CI: [ 0.25 ,  0.401 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs All-OpenAI
## ----------------------------------------
## Proportions:  0.628  vs  0.495
## Difference:  0.133
## Chi-squared:  29.292
## Degrees of freedom:  1
## P-value:  0.00000006225
## 95% CI: [ 0.085 ,  0.18 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.628  vs  0.444
## Difference:  0.183
## Chi-squared:  53.189
## Degrees of freedom:  1
## P-value:  0.0000000000003029
## 95% CI: [ 0.134 ,  0.232 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs All-Gemini
## ----------------------------------------
## Proportions:  0.628  vs  0.396
## Difference:  0.231
## Chi-squared:  72.959
## Degrees of freedom:  1
## P-value:  0.00000000000000001324
## 95% CI: [ 0.179 ,  0.284 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.404  vs  0.397
## Difference:  0.007
## Chi-squared:  0.021
## Degrees of freedom:  1
## P-value:  0.8835
## 95% CI: [ -0.057 ,  0.071 ]
## Significant:  NO
##
##  GPT-4.1 vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.404  vs  0.417
## Difference:  -0.014
## Chi-squared:  0.132
## Degrees of freedom:  1
## P-value:  0.7164
## 95% CI: [ -0.078 ,  0.051 ]
## Significant:  NO
##
##  GPT-4.1 vs o4-mini
## ----------------------------------------
## Proportions:  0.404  vs  0.501
## Difference:  -0.098
## Chi-squared:  11.833
## Degrees of freedom:  1
## P-value:  0.0005821
## 95% CI: [ -0.153 ,  -0.042 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.404  vs  0.445
## Difference:  -0.042
## Chi-squared:  1.527
## Degrees of freedom:  1
## P-value:  0.2166
## 95% CI: [ -0.106 ,  0.023 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.404  vs  0.394
## Difference:  0.01
## Chi-squared:  0.059
## Degrees of freedom:  1
## P-value:  0.8073
## 95% CI: [ -0.054 ,  0.074 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.404  vs  0.302
## Difference:  0.102
## Chi-squared:  6.713
## Degrees of freedom:  1
## P-value:  0.009572
## 95% CI: [ 0.026 ,  0.178 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs All-OpenAI
## ----------------------------------------
## Proportions:  0.404  vs  0.495
## Difference:  -0.091
## Chi-squared:  13.708
## Degrees of freedom:  1
## P-value:  0.0002135
## 95% CI: [ -0.139 ,  -0.043 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.404  vs  0.444
## Difference:  -0.04
## Chi-squared:  2.49
## Degrees of freedom:  1
## P-value:  0.1146
## 95% CI: [ -0.09 ,  0.009 ]
## Significant:  NO
```

```
##
##  GPT-4.1 vs All-Gemini
## ----------------------------------------
## Proportions:  0.404  vs  0.396
## Difference:  0.008
## Chi-squared:  0.056
## Degrees of freedom:  1
## P-value:  0.8129
## 95% CI: [ -0.046 ,  0.061 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.397  vs  0.417
## Difference:  -0.02
## Chi-squared:  0.331
## Degrees of freedom:  1
## P-value:  0.5649
## 95% CI: [ -0.085 ,  0.044 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs o4-mini
## ----------------------------------------
## Proportions:  0.397  vs  0.501
## Difference:  -0.104
## Chi-squared:  13.555
## Degrees of freedom:  1
## P-value:  0.0002317
## 95% CI: [ -0.16 ,  -0.049 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.397  vs  0.445
## Difference:  -0.048
## Chi-squared:  2.096
## Degrees of freedom:  1
## P-value:  0.1477
## 95% CI: [ -0.113 ,  0.016 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.397  vs  0.394
## Difference:  0.003
## Chi-squared:  0.001
## Degrees of freedom:  1
## P-value:  0.9749
## 95% CI: [ -0.061 ,  0.067 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.397  vs  0.302
## Difference:  0.095
## Chi-squared:  5.862
## Degrees of freedom:  1
## P-value:  0.01547
## 95% CI: [ 0.019 ,  0.171 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs All-OpenAI
## ----------------------------------------
## Proportions:  0.397  vs  0.495
## Difference:  -0.098
## Chi-squared:  15.835
## Degrees of freedom:  1
## P-value:  0.00006912
## 95% CI: [ -0.146 ,  -0.05 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.397  vs  0.444
## Difference:  -0.047
## Chi-squared:  3.42
## Degrees of freedom:  1
## P-value:  0.06442
## 95% CI: [ -0.096 ,  0.002 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs All-Gemini
## ----------------------------------------
## Proportions:  0.397  vs  0.396
## Difference:  0.001
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.052 ,  0.054 ]
## Significant:  NO
##
##  ChatGPT-4o vs o4-mini
## ----------------------------------------
## Proportions:  0.417  vs  0.501
## Difference:  -0.084
## Chi-squared:  8.703
## Degrees of freedom:  1
## P-value:  0.003178
## 95% CI: [ -0.14 ,  -0.028 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.417  vs  0.445
## Difference:  -0.028
## Chi-squared:  0.651
## Degrees of freedom:  1
## P-value:  0.4196
## 95% CI: [ -0.093 ,  0.037 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.417  vs  0.394
## Difference:  0.023
## Chi-squared:  0.453
## Degrees of freedom:  1
## P-value:  0.501
## 95% CI: [ -0.041 ,  0.088 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.417  vs  0.302
## Difference:  0.116
## Chi-squared:  8.605
## Degrees of freedom:  1
## P-value:  0.003352
## 95% CI: [ 0.04 ,  0.192 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs All-OpenAI
## ----------------------------------------
## Proportions:  0.417  vs  0.495
## Difference:  -0.077
## Chi-squared:  9.868
## Degrees of freedom:  1
## P-value:  0.001682
## 95% CI: [ -0.126 ,  -0.029 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.417  vs  0.444
## Difference:  -0.027
## Chi-squared:  1.057
## Degrees of freedom:  1
## P-value:  0.3038
## 95% CI: [ -0.076 ,  0.023 ]
## Significant:  NO
##
##  ChatGPT-4o vs All-Gemini
## ----------------------------------------
## Proportions:  0.417  vs  0.396
## Difference:  0.021
## Chi-squared:  0.563
## Degrees of freedom:  1
## P-value:  0.4529
## 95% CI: [ -0.032 ,  0.075 ]
## Significant:  NO
##
##  o4-mini vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.501  vs  0.445
## Difference:  0.056
## Chi-squared:  3.8
## Degrees of freedom:  1
## P-value:  0.05126
## 95% CI: [ 0 ,  0.112 ]
## Significant:  NO
##
##  o4-mini vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.501  vs  0.394
## Difference:  0.107
## Chi-squared:  14.383
## Degrees of freedom:  1
## P-value:  0.0001492
## 95% CI: [ 0.052 ,  0.163 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.501  vs  0.302
## Difference:  0.2
## Chi-squared:  29.961
## Degrees of freedom:  1
## P-value:  0.00000004408
## 95% CI: [ 0.131 ,  0.268 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs All-OpenAI
## ----------------------------------------
## Proportions:  0.501  vs  0.495
## Difference:  0.006
## Chi-squared:  0.101
## Degrees of freedom:  1
## P-value:  0.7506
## 95% CI: [ -0.03 ,  0.043 ]
## Significant:  NO
##
##  o4-mini vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.501  vs  0.444
## Difference:  0.057
## Chi-squared:  8.781
## Degrees of freedom:  1
## P-value:  0.003045
## 95% CI: [ 0.019 ,  0.095 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs All-Gemini
## ----------------------------------------
## Proportions:  0.501  vs  0.396
## Difference:  0.105
## Chi-squared:  23.514
## Degrees of freedom:  1
## P-value:  0.00000124
## 95% CI: [ 0.062 ,  0.148 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.445  vs  0.394
## Difference:  0.051
## Chi-squared:  2.387
## Degrees of freedom:  1
## P-value:  0.1224
## 95% CI: [ -0.013 ,  0.116 ]
## Significant:  NO
##
##  Gemini-2.5 vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.445  vs  0.302
## Difference:  0.144
## Chi-squared:  13.168
## Degrees of freedom:  1
## P-value:  0.0002848
## 95% CI: [ 0.067 ,  0.22 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5 vs All-OpenAI
## ----------------------------------------
## Proportions:  0.445  vs  0.495
## Difference:  -0.05
## Chi-squared:  3.971
## Degrees of freedom:  1
## P-value:  0.0463
## 95% CI: [ -0.098 ,  -0.001 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5 vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.445  vs  0.444
## Difference:  0.001
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.049 ,  0.051 ]
## Significant:  NO
##
##  Gemini-2.5 vs All-Gemini
## ----------------------------------------
## Proportions:  0.445  vs  0.396
## Difference:  0.049
## Chi-squared:  3.233
## Degrees of freedom:  1
## P-value:  0.07216
## 95% CI: [ -0.005 ,  0.103 ]
## Significant:  NO
```

```
##
##  Gemini-2.0-Flash vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.394  vs  0.302
## Difference:  0.092
## Chi-squared:  5.49
## Degrees of freedom:  1
## P-value:  0.01912
## 95% CI: [ 0.016 ,  0.168 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs All-OpenAI
## ----------------------------------------
## Proportions:  0.394  vs  0.495
## Difference:  -0.101
## Chi-squared:  16.859
## Degrees of freedom:  1
## P-value:  0.00004025
## 95% CI: [ -0.149 ,  -0.053 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.394  vs  0.444
## Difference:  -0.05
## Chi-squared:  3.895
## Degrees of freedom:  1
## P-value:  0.04843
## 95% CI: [ -0.099 ,  -0.001 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs All-Gemini
## ----------------------------------------
## Proportions:  0.394  vs  0.396
## Difference:  -0.002
## Chi-squared:  0.001
## Degrees of freedom:  1
## P-value:  0.9813
## 95% CI: [ -0.055 ,  0.051 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-GPT-Image vs All-OpenAI
## ----------------------------------------
## Proportions:  0.302  vs  0.495
## Difference:  -0.193
## Chi-squared:  32.829
## Degrees of freedom:  1
## P-value:  0.00000001006
## 95% CI: [ -0.256 ,  -0.131 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-GPT-Image vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.302  vs  0.444
## Difference:  -0.142
## Chi-squared:  17.461
## Degrees of freedom:  1
## P-value:  0.00002932
## 95% CI: [ -0.206 ,  -0.079 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-GPT-Image vs All-Gemini
## ----------------------------------------
## Proportions:  0.302  vs  0.396
## Difference:  -0.094
## Chi-squared:  7.15
## Degrees of freedom:  1
## P-value:  0.007497
## 95% CI: [ -0.161 ,  -0.027 ]
## Significant:  YES (p < 0.05)
##
##  All-OpenAI vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.495  vs  0.444
## Difference:  0.051
## Chi-squared:  14.642
## Degrees of freedom:  1
## P-value:  0.00013
## 95% CI: [ 0.025 ,  0.077 ]
## Significant:  YES (p < 0.05)
##
##  All-OpenAI vs All-Gemini
## ----------------------------------------
## Proportions:  0.495  vs  0.396
## Difference:  0.099
## Chi-squared:  34.875
## Degrees of freedom:  1
## P-value:  0.000000003516
## 95% CI: [ 0.066 ,  0.131 ]
## Significant:  YES (p < 0.05)
##
##  Other-OpenAI vs All-Gemini
## ----------------------------------------
## Proportions:  0.444  vs  0.396
## Difference:  0.048
## Chi-squared:  7.366
## Degrees of freedom:  1
## P-value:  0.006647
## 95% CI: [ 0.013 ,  0.083 ]
## Significant:  YES (p < 0.05)
```

```r
# Summary table
novel_48_summary <- novel_48_results %>%
  select(comparison, diff, chi_squared, p_value, significant) %>%
  mutate(diff = round(diff, 3),
         p_value = round(p_value, 4))

cat("\n\nSummary Table - 48 Novel Tasks:\n")
```

```
##
##
## Summary Table - 48 Novel Tasks:
```

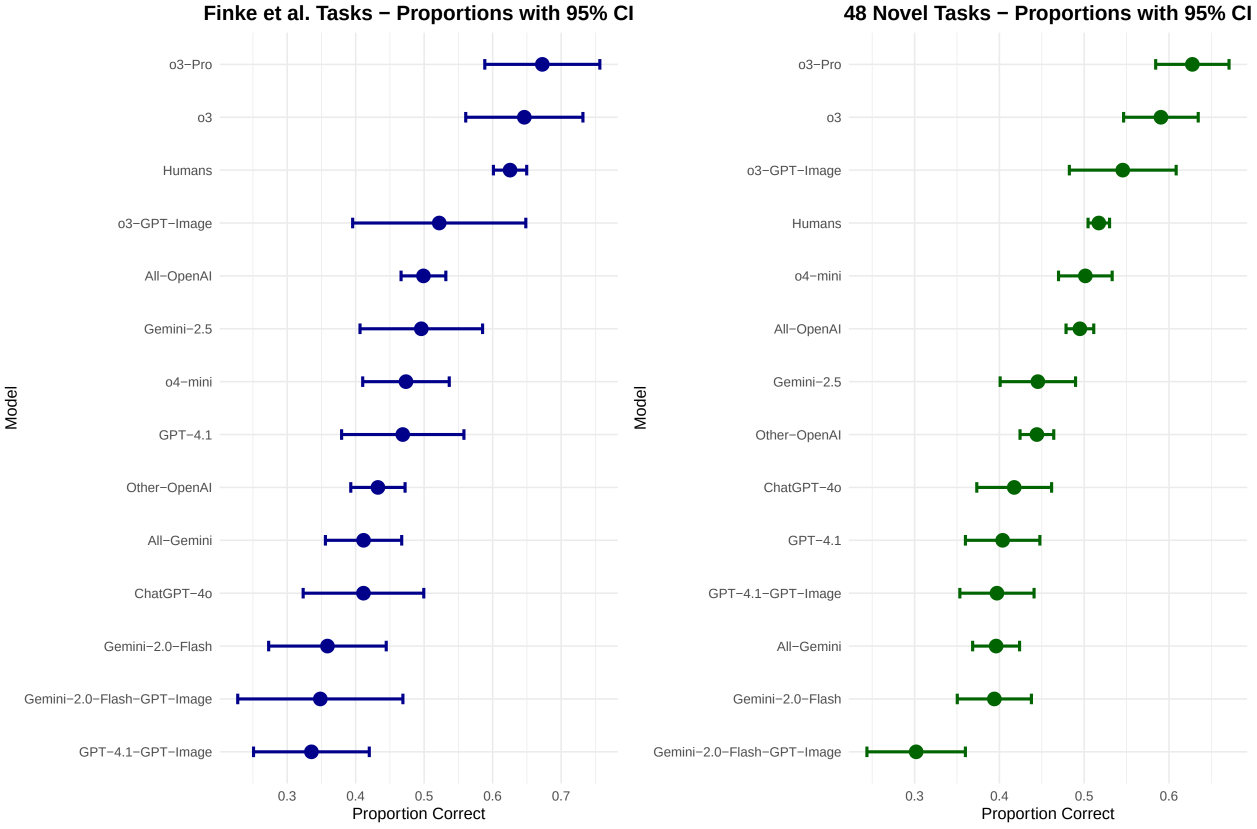
```
print(kable(novel_48_summary, format = "simple"))
```

```
##
##
##                 comparison                                              diff   chi_squared    p_value   sig
## ------------   ----------------------------------------------   -------   ------------   --------   ---
## X-squared      Humans vs o3                                          -0.073     9.2646612    0.0023   TRU
## X-squared1     Humans vs o3-GPT-Image                                -0.028     0.6290016    0.4277   FAL
## X-squared2     Humans vs o3-Pro                                      -0.110    21.2469930    0.0000   TRU
## X-squared3     Humans vs GPT-4.1                                      0.113    22.4091749    0.0000   TRU
## X-squared4     Humans vs GPT-4.1-GPT-Image                            0.120    25.1694089    0.0000   TRU
## X-squared5     Humans vs ChatGPT-4o                                   0.100    17.3032108    0.0000   TRU
## X-squared6     Humans vs o4-mini                                      0.016     0.7727878    0.3794   FAL
## X-squared7     Humans vs Gemini-2.5                                   0.072     8.9054240    0.0028   TRU
## X-squared8     Humans vs Gemini-2.0-Flash                             0.123    26.4868070    0.0000   TRU
## X-squared9     Humans vs Gemini-2.0-Flash-GPT-Image                   0.215    41.9801273    0.0000   TRU
## X-squared10    Humans vs All-OpenAI                                   0.022     4.3843160    0.0363   TRU
## X-squared11    Humans vs Other-OpenAI                                 0.073    36.2063594    0.0000   TRU
## X-squared12    Humans vs All-Gemini                                   0.121    58.1419995    0.0000   TRU
## X-squared13    o3 vs o3-GPT-Image                                     0.045     1.1484397    0.2839   FAL
## X-squared14    o3 vs o3-Pro                                          -0.037     1.2340252    0.2666   FAL
## X-squared15    o3 vs GPT-4.1                                          0.187    32.7074528    0.0000   TRU
## X-squared16    o3 vs GPT-4.1-GPT-Image                                0.193    35.1352676    0.0000   TRU
## X-squared17    o3 vs ChatGPT-4o                                       0.173    28.0591389    0.0000   TRU
## X-squared18    o3 vs o4-mini                                          0.089     9.8553086    0.0017   TRU
## X-squared19    o3 vs Gemini-2.5                                       0.145    19.6756829    0.0000   TRU
## X-squared20    o3 vs Gemini-2.0-Flash                                 0.196    36.2771737    0.0000   TRU
## X-squared21    o3 vs Gemini-2.0-Flash-GPT-Image                       0.289    52.1900627    0.0000   TRU
## X-squared22    o3 vs All-OpenAI                                       0.096    15.0962780    0.0001   TRU
## X-squared23    o3 vs Other-OpenAI                                     0.146    33.7832552    0.0000   TRU
## X-squared24    o3 vs All-Gemini                                       0.194    51.5227679    0.0000   TRU
## X-squared25    o3-GPT-Image vs o3-Pro                                -0.082     4.1560028    0.0415   TRU
## X-squared26    o3-GPT-Image vs GPT-4.1                                0.142    12.4019415    0.0004   TRU
## X-squared27    o3-GPT-Image vs GPT-4.1-GPT-Image                      0.148    13.6588637    0.0002   TRU
## X-squared28    o3-GPT-Image vs ChatGPT-4o                             0.128    10.0503677    0.0015   TRU
## X-squared29    o3-GPT-Image vs o4-mini                                0.044     1.3254849    0.2496   FAL
## X-squared30    o3-GPT-Image vs Gemini-2.5                             0.100     6.0362491    0.0140   TRU
## X-squared31    o3-GPT-Image vs Gemini-2.0-Flash                       0.151    14.2562529    0.0002   TRU
## X-squared32    o3-GPT-Image vs Gemini-2.0-Flash-GPT-Image             0.244    28.1908104    0.0000   TRU
## X-squared33    o3-GPT-Image vs All-OpenAI                             0.051     2.1039117    0.1469   FAL
## X-squared34    o3-GPT-Image vs Other-OpenAI                           0.101     8.6249596    0.0033   TRU
## X-squared35    o3-GPT-Image vs All-Gemini                             0.149    17.6998569    0.0000   TRU
## X-squared36    o3-Pro vs GPT-4.1                                      0.224    47.2109517    0.0000   TRU
## X-squared37    o3-Pro vs GPT-4.1-GPT-Image                            0.230    50.0964716    0.0000   TRU
## X-squared38    o3-Pro vs ChatGPT-4o                                   0.210    41.6301913    0.0000   TRU
## X-squared39    o3-Pro vs o4-mini                                      0.126    20.0394952    0.0000   TRU
## X-squared40    o3-Pro vs Gemini-2.5                                   0.182    31.3203273    0.0000   TRU
## X-squared41    o3-Pro vs Gemini-2.0-Flash                             0.234    51.4475032    0.0000   TRU
## X-squared42    o3-Pro vs Gemini-2.0-Flash-GPT-Image                   0.326    66.7094603    0.0000   TRU
## X-squared43    o3-Pro vs All-OpenAI                                   0.133    29.2922114    0.0000   TRU
## X-squared44    o3-Pro vs Other-OpenAI                                 0.183    53.1890793    0.0000   TRU
## X-squared45   o3-Pro vs All-Gemini                              0.231    72.9592823    0.0000  TRU
## X-squared46   GPT-4.1 vs GPT-4.1-GPT-Image                      0.007     0.0214554    0.8835  FAL
## X-squared47   GPT-4.1 vs ChatGPT-4o                            -0.014     0.1320088    0.7164  FAL
## X-squared48   GPT-4.1 vs o4-mini                               -0.098    11.8325063    0.0006  TRU
## X-squared49   GPT-4.1 vs Gemini-2.5                            -0.042     1.5268506    0.2166  FAL
## X-squared50   GPT-4.1 vs Gemini-2.0-Flash                       0.010     0.0594708    0.8073  FAL
## X-squared51   GPT-4.1 vs Gemini-2.0-Flash-GPT-Image             0.102     6.7128366    0.0096  TRU
## X-squared52   GPT-4.1 vs All-OpenAI                            -0.091    13.7084949    0.0002  TRU
## X-squared53   GPT-4.1 vs Other-OpenAI                          -0.040     2.4899915    0.1146  FAL
## X-squared54   GPT-4.1 vs All-Gemini                             0.008     0.0560061    0.8129  FAL
## X-squared55   GPT-4.1-GPT-Image vs ChatGPT-4o                  -0.020     0.3312666    0.5649  FAL
## X-squared56   GPT-4.1-GPT-Image vs o4-mini                     -0.104    13.5546564    0.0002  TRU
## X-squared57   GPT-4.1-GPT-Image vs Gemini-2.5                  -0.048     2.0958236    0.1477  FAL
## X-squared58   GPT-4.1-GPT-Image vs Gemini-2.0-Flash             0.003     0.0009879    0.9749  FAL
## X-squared59   GPT-4.1-GPT-Image vs Gemini-2.0-Flash-GPT-Image   0.095     5.8617428    0.0155  TRU
## X-squared60   GPT-4.1-GPT-Image vs All-OpenAI                  -0.098    15.8346639    0.0001  TRU
## X-squared61   GPT-4.1-GPT-Image vs Other-OpenAI                -0.047     3.4198201    0.0644  FAL
## X-squared62   GPT-4.1-GPT-Image vs All-Gemini                   0.001     0.0000000    1.0000  FAL
## X-squared63   ChatGPT-4o vs o4-mini                            -0.084     8.7026171    0.0032  TRU
## X-squared64   ChatGPT-4o vs Gemini-2.5                         -0.028     0.6514510    0.4196  FAL
## X-squared65   ChatGPT-4o vs Gemini-2.0-Flash                    0.023     0.4528312    0.5010  FAL
## X-squared66   ChatGPT-4o vs Gemini-2.0-Flash-GPT-Image          0.116     8.6053270    0.0034  TRU
## X-squared67   ChatGPT-4o vs All-OpenAI                         -0.077     9.8676401    0.0017  TRU
## X-squared68   ChatGPT-4o vs Other-OpenAI                       -0.027     1.0572791    0.3038  FAL
## X-squared69   ChatGPT-4o vs All-Gemini                          0.021     0.5634439    0.4529  FAL
## X-squared70   o4-mini vs Gemini-2.5                             0.056     3.7997614    0.0513  FAL
## X-squared71   o4-mini vs Gemini-2.0-Flash                       0.107    14.3828219    0.0001  TRU
## X-squared72   o4-mini vs Gemini-2.0-Flash-GPT-Image             0.200    29.9612588    0.0000  TRU
## X-squared73   o4-mini vs All-OpenAI                             0.006     0.1010169    0.7506  FAL
## X-squared74   o4-mini vs Other-OpenAI                           0.057     8.7805714    0.0030  TRU
## X-squared75   o4-mini vs All-Gemini                             0.105    23.5141392    0.0000  TRU
## X-squared76   Gemini-2.5 vs Gemini-2.0-Flash                    0.051     2.3869120    0.1224  FAL
## X-squared77   Gemini-2.5 vs Gemini-2.0-Flash-GPT-Image          0.144    13.1675437    0.0003  TRU
## X-squared78   Gemini-2.5 vs All-OpenAI                         -0.050     3.9707517    0.0463  TRU
## X-squared79   Gemini-2.5 vs Other-OpenAI                        0.001     0.0000000    1.0000  FAL
## X-squared80   Gemini-2.5 vs All-Gemini                          0.049     3.2331996    0.0722  FAL
## X-squared81   Gemini-2.0-Flash vs Gemini-2.0-Flash-GPT-Image    0.092     5.4902583    0.0191  TRU
## X-squared82   Gemini-2.0-Flash vs All-OpenAI                   -0.101    16.8594895    0.0000  TRU
## X-squared83   Gemini-2.0-Flash vs Other-OpenAI                 -0.050     3.8949474    0.0484  TRU
## X-squared84   Gemini-2.0-Flash vs All-Gemini                   -0.002     0.0005478    0.9813  FAL
## X-squared85   Gemini-2.0-Flash-GPT-Image vs All-OpenAI         -0.193    32.8291727    0.0000  TRU
## X-squared86   Gemini-2.0-Flash-GPT-Image vs Other-OpenAI       -0.142    17.4614037    0.0000  TRU
## X-squared87   Gemini-2.0-Flash-GPT-Image vs All-Gemini         -0.094     7.1498590    0.0075  TRU
## X-squared88   All-OpenAI vs Other-OpenAI                        0.051    14.6417706    0.0001  TRU
## X-squared89   All-OpenAI vs All-Gemini                          0.099    34.8746299    0.0000  TRU
## X-squared90   Other-OpenAI vs All-Gemini                        0.048     7.3658646    0.0066  TRU
```

Visualization of All Comparisons

```
# Plot 1: Proportions with confidence intervals for Finke tasks
finke_plot <- ggplot(finke_data, aes(x = reorder(model, proportion), y = proportion)) +
  geom_point(size = 4, color = "darkblue") +
  geom_errorbar(aes(ymin = proportion - 1.96 * sqrt(proportion * (1 - proportion) / total),
                    ymax = proportion + 1.96 * sqrt(proportion * (1 - proportion) / total)),
```

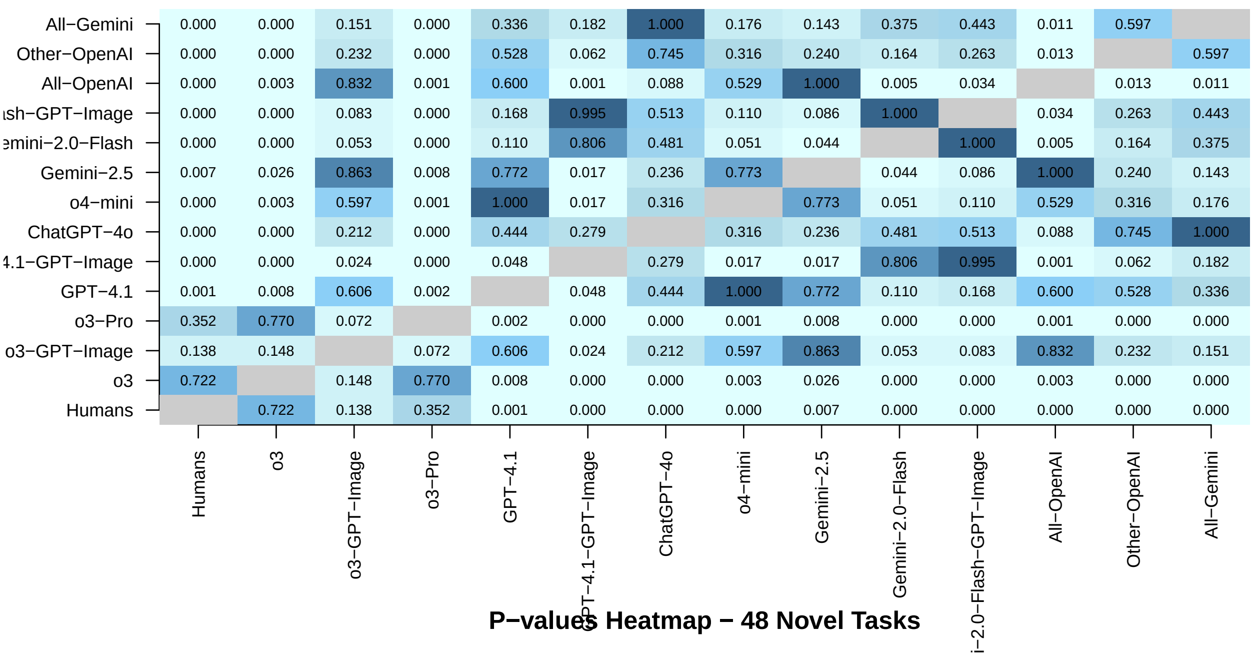
```r
                width = 0.2, size = 1, color = "darkblue") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Finke et al. Tasks - Proportions with 95% CI",
       x = "Model",
       y = "Proportion Correct") +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))

# Plot 2: Proportions with confidence intervals for 48 Novel tasks
novel_48_plot <- ggplot(novel_data, aes(x = reorder(model, proportion), y = proportion)) +
  geom_point(size = 4, color = "darkgreen") +
  geom_errorbar(aes(ymin = proportion - 1.96 * sqrt(proportion * (1 - proportion) / total),
                    ymax = proportion + 1.96 * sqrt(proportion * (1 - proportion) / total)),
                width = 0.2, size = 1, color = "darkgreen") +
  coord_flip() +
  theme_minimal() +
  labs(title = "48 Novel Tasks - Proportions with 95% CI",
       x = "Model",
       y = "Proportion Correct") +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))

# Combine plots
combined_plot <- finke_plot + novel_48_plot
print(combined_plot)
```



Heatmap of P-values

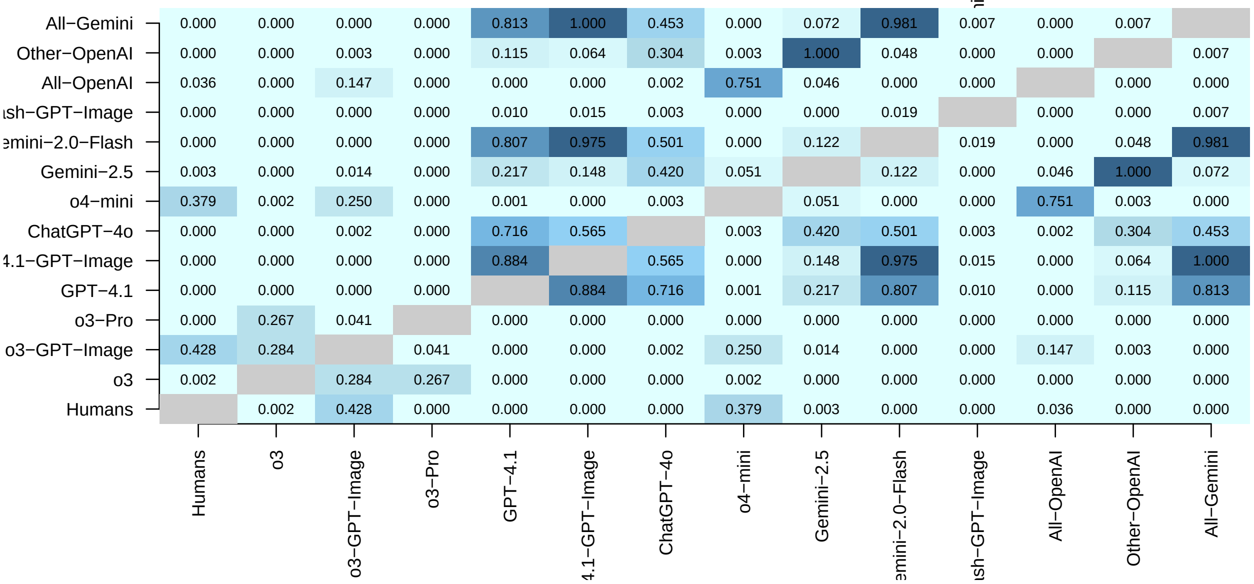
```r
# Create matrix of p-values for Finke tasks
finke_models <- finke_data$model
finke_pval_matrix <- matrix(NA, nrow = length(finke_models), ncol = length(finke_models))
rownames(finke_pval_matrix) <- finke_models
colnames(finke_pval_matrix) <- finke_models

for (i in 1:nrow(finke_results)) {
  row_idx <- which(finke_models == finke_results$model1[i])
  col_idx <- which(finke_models == finke_results$model2[i])
  finke_pval_matrix[row_idx, col_idx] <- finke_results$p_value[i]
  finke_pval_matrix[col_idx, row_idx] <- finke_results$p_value[i]
}

# Set diagonal to NA
diag(finke_pval_matrix) <- NA

# Create matrix of p-values for 48 Novel tasks
novel_models <- novel_data$model
novel_pval_matrix <- matrix(NA, nrow = length(novel_models), ncol = length(novel_models))
rownames(novel_pval_matrix) <- novel_models
colnames(novel_pval_matrix) <- novel_models

for (i in 1:nrow(novel_48_results)) {
  row_idx <- which(novel_models == novel_48_results$model1[i])
  col_idx <- which(novel_models == novel_48_results$model2[i])
  novel_pval_matrix[row_idx, col_idx] <- novel_48_results$p_value[i]
  novel_pval_matrix[col_idx, row_idx] <- novel_48_results$p_value[i]
}

# Set diagonal to NA
diag(novel_pval_matrix) <- NA

# Plot heatmaps
par(mfrow = c(2, 1), mar = c(6, 6, 3, 2))  # Increase margins for labels

# Define color palette
col_palette <- colorRampPalette(c("lightcyan", "lightblue", "lightskyblue", "steelblue4"))(20)

# Finke heatmap
image(finke_pval_matrix, axes = FALSE, col = col_palette, main = "P-values Heatmap - Finke Tasks")
axis(1, at = seq(0, 1, length.out = length(finke_models)), labels = finke_models,
     las = 2, cex.axis = 0.8)  # las=2 makes labels perpendicular, cex.axis makes them smaller
axis(2, at = seq(0, 1, length.out = length(finke_models)), labels = finke_models,
     las = 2, cex.axis = 0.8)

# Add gray color for diagonal
for (i in 1:length(finke_models)) {
  x_pos <- (i - 1) / (length(finke_models) - 1)
  y_pos <- (i - 1) / (length(finke_models) - 1)
  rect(x_pos - 0.5 / (length(finke_models) - 1), y_pos - 0.5 / (length(finke_models) - 1),
       x_pos + 0.5 / (length(finke_models) - 1), y_pos + 0.5 / (length(finke_models) - 1),
       col = "gray80", border = NA)
}

# Add p-values to the plot
for (i in 1:nrow(finke_pval_matrix)) {
  for (j in 1:ncol(finke_pval_matrix)) {
    if (!is.na(finke_pval_matrix[i, j])) {
      x_pos <- (j - 1) / (ncol(finke_pval_matrix) - 1)
      y_pos <- (i - 1) / (nrow(finke_pval_matrix) - 1)
      text(x_pos, y_pos, sprintf("%.3f", finke_pval_matrix[i, j]), cex = 0.7)
    }
  }
}

# 48 Novel heatmap
image(novel_pval_matrix, axes = FALSE, col = col_palette, main = "P-values Heatmap - 48 Novel Tasks")
axis(1, at = seq(0, 1, length.out = length(novel_models)), labels = novel_models,
     las = 2, cex.axis = 0.8)  # las=2 makes labels perpendicular, cex.axis makes them smaller
axis(2, at = seq(0, 1, length.out = length(novel_models)), labels = novel_models,
     las = 2, cex.axis = 0.8)

# Add gray color for diagonal
for (i in 1:length(novel_models)) {
  x_pos <- (i - 1) / (length(novel_models) - 1)
  y_pos <- (i - 1) / (length(novel_models) - 1)
  rect(x_pos - 0.5 / (length(novel_models) - 1), y_pos - 0.5 / (length(novel_models) - 1),
       x_pos + 0.5 / (length(novel_models) - 1), y_pos + 0.5 / (length(novel_models) - 1),
       col = "gray80", border = NA)
}

# Add p-values to the plot
for (i in 1:nrow(novel_pval_matrix)) {
  for (j in 1:ncol(novel_pval_matrix)) {
    if (!is.na(novel_pval_matrix[i, j])) {
      x_pos <- (j - 1) / (ncol(novel_pval_matrix) - 1)
      y_pos <- (i - 1) / (nrow(novel_pval_matrix) - 1)
      text(x_pos, y_pos, sprintf("%.3f", novel_pval_matrix[i, j]), cex = 0.7)
    }
  }
}
```

P−values Heatmap − Finke Tasks



	Humans	o3	o3−GPT−Image	o3−Pro	GPT−4.1	PT−4.1−GPT−Image	ChatGPT−4o	o4−mini	Gemini−2.5	Gemini−2.0−Flash	i−2.0−Flash−GPT−Image	All−OpenAI	Other−OpenAI	All−Gemini
All−Gemini	0.000	0.000	0.151	0.000	0.336	0.182	1.000	0.176	0.143	0.375	0.443	0.011	0.597	
Other−OpenAI	0.000	0.000	0.232	0.000	0.528	0.062	0.745	0.316	0.240	0.164	0.263	0.013		0.597
All−OpenAI	0.000	0.003	0.832	0.001	0.600	0.001	0.088	0.529	1.000	0.005	0.034		0.013	0.011
ısh−GPT−Image	0.000	0.000	0.083	0.000	0.168	0.995	0.513	0.110	0.086	1.000		0.034	0.263	0.443
əmini−2.0−Flash	0.000	0.000	0.053	0.000	0.110	0.806	0.481	0.051	0.044		1.000	0.005	0.164	0.375
Gemini−2.5	0.007	0.026	0.863	0.008	0.772	0.017	0.236	0.773		0.044	0.086	1.000	0.240	0.143
o4−mini	0.000	0.003	0.597	0.001	1.000	0.017	0.316		0.773	0.051	0.110	0.529	0.316	0.176
ChatGPT−4o	0.000	0.000	0.212	0.000	0.444	0.279		0.316	0.236	0.481	0.513	0.088	0.745	1.000
4.1−GPT−Image	0.000	0.000	0.024	0.000	0.048		0.279	0.017	0.017	0.806	0.995	0.001	0.062	0.182
GPT−4.1	0.001	0.008	0.606	0.002		0.048	0.444	1.000	0.772	0.110	0.168	0.600	0.528	0.336
o3−Pro	0.352	0.770	0.072		0.002	0.000	0.000	0.001	0.008	0.000	0.000	0.001	0.000	0.000
o3−GPT−Image	0.138	0.148		0.072	0.606	0.024	0.212	0.597	0.863	0.053	0.083	0.832	0.232	0.151
o3	0.722		0.148	0.770	0.008	0.000	0.000	0.003	0.026	0.000	0.000	0.003	0.000	0.000
Humans		0.722	0.138	0.352	0.001	0.000	0.000	0.000	0.007	0.000	0.000	0.000	0.000	0.000

P−values Heatmap − 48 Novel Tasks



	Humans	o3	o3−GPT−Image	o3−Pro	GPT−4.1	4.1−GPT−Image	ChatGPT−4o	o4−mini	Gemini−2.5	əmini−2.0−Flash	ısh−GPT−Image	All−OpenAI	Other−OpenAI	All−Gemini
All−Gemini	0.000	0.000	0.000	0.000	0.813	1.000	0.453	0.000	0.072	0.981	0.007	0.000	0.007	
Other−OpenAI	0.000	0.000	0.003	0.000	0.115	0.064	0.304	0.003	1.000	0.048	0.000	0.000		0.007
All−OpenAI	0.036	0.000	0.147	0.000	0.000	0.000	0.002	0.751	0.046	0.000	0.000		0.000	0.000
ısh−GPT−Image	0.000	0.000	0.000	0.000	0.010	0.015	0.003	0.000	0.000	0.019		0.000	0.000	0.007
əmini−2.0−Flash	0.000	0.000	0.000	0.000	0.807	0.975	0.501	0.000	0.122		0.019	0.000	0.048	0.981
Gemini−2.5	0.003	0.000	0.014	0.000	0.217	0.148	0.420	0.051		0.122	0.000	0.046	1.000	0.072
o4−mini	0.379	0.002	0.250	0.000	0.001	0.000	0.003		0.051	0.000	0.000	0.751	0.003	0.000
ChatGPT−4o	0.000	0.000	0.002	0.000	0.716	0.565		0.003	0.420	0.501	0.003	0.002	0.304	0.453
4.1−GPT−Image	0.000	0.000	0.000	0.000	0.884		0.565	0.000	0.148	0.975	0.015	0.000	0.064	1.000
GPT−4.1	0.000	0.000	0.000	0.000		0.884	0.716	0.001	0.217	0.807	0.010	0.000	0.115	0.813
o3−Pro	0.000	0.267	0.041		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
o3−GPT−Image	0.428	0.284		0.041	0.000	0.000	0.002	0.250	0.014	0.000	0.000	0.147	0.003	0.000
o3	0.002		0.284	0.267	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000
Humans		0.002	0.428	0.000	0.000	0.000	0.000	0.379	0.003	0.000	0.000	0.036	0.000	0.000

Summary of Significant Differences

```r
# Count significant differences for each task
finke_sig_count <- sum(finke_results$significant)
novel_48_sig_count <- sum(novel_48_results$significant)

cat("Summary of Significant Differences:\n")
```

```
## Summary of Significant Differences:
```

```r
cat(paste(rep("=", 50), collapse = ""), "\n")
```

```
## ==================================================
```

```r
cat("Finke Tasks:\n")
```

```
## Finke Tasks:
```

```r
cat("  Total comparisons:", nrow(finke_results), "\n")
```

```
##   Total comparisons: 91
```

```r
cat("  Significant differences:", finke_sig_count, "\n")
```

```
##   Significant differences: 40
```

```r
cat("  Percentage significant:", round(finke_sig_count / nrow(finke_results) * 100, 1), "%\n\n")
```

```
##   Percentage significant: 44 %
```

```r
cat("48 Novel Tasks:\n")
```

```
## 48 Novel Tasks:
```

```r
cat("  Total comparisons:", nrow(novel_48_results), "\n")
```

```
##   Total comparisons: 91
```

```r
cat("  Significant differences:", novel_48_sig_count, "\n")
```

```
##   Significant differences: 64
```

```r
cat("  Percentage significant:", round(novel_48_sig_count / nrow(novel_48_results) * 100, 1), "%\n\n")
```

```
##   Percentage significant: 70.3 %
```

```r
# Show which comparisons are significant
cat("Significant Comparisons in Finke Tasks:\n")
```

```
## Significant Comparisons in Finke Tasks:
```

```r
finke_sig <- finke_results[finke_results$significant, c("comparison", "diff", "p_value")]
if (nrow(finke_sig) > 0) {
  print(kable(finke_sig, format = "simple", digits = 4))
} else {
  cat("  None\n")
}
```

```
##
##
##               comparison                                     diff   p_value
## ------------  -----------------------------------------  --------  --------
## X-squared3    Humans vs GPT-4.1                            0.1566    0.0010
## X-squared4    Humans vs GPT-4.1-GPT-Image                  0.2900    0.0000
## X-squared5    Humans vs ChatGPT-4o                         0.2140    0.0000
## X-squared6    Humans vs o4-mini                            0.1519    0.0000
## X-squared7    Humans vs Gemini-2.5                         0.1296    0.0067
## X-squared8    Humans vs Gemini-2.0-Flash                   0.2666    0.0000
## X-squared9    Humans vs Gemini-2.0-Flash-GPT-Image         0.2770    0.0000
## X-squared10   Humans vs All-OpenAI                         0.1264    0.0000
## X-squared11   Humans vs Other-OpenAI                       0.1929    0.0000
## X-squared12   Humans vs All-Gemini                         0.2139    0.0000
## X-squared15   o3 vs GPT-4.1                                0.1775    0.0083
## X-squared16   o3 vs GPT-4.1-GPT-Image                      0.3108    0.0000
## X-squared17   o3 vs ChatGPT-4o                             0.2348    0.0004
## X-squared18   o3 vs o4-mini                                0.1728    0.0028
## X-squared19   o3 vs Gemini-2.5                             0.1504    0.0262
## X-squared20   o3 vs Gemini-2.0-Flash                       0.2874    0.0000
## X-squared21   o3 vs Gemini-2.0-Flash-GPT-Image             0.2978    0.0003
## X-squared22   o3 vs All-OpenAI                             0.1473    0.0033
## X-squared23   o3 vs Other-OpenAI                           0.2137    0.0000
## X-squared24   o3 vs All-Gemini                             0.2347    0.0000
## X-squared27   o3-GPT-Image vs GPT-4.1-GPT-Image            0.1866    0.0245
## X-squared36   o3-Pro vs GPT-4.1                            0.2037    0.0022
## X-squared37   o3-Pro vs GPT-4.1-GPT-Image                  0.3371    0.0000
## X-squared38   o3-Pro vs ChatGPT-4o                         0.2611    0.0001
## X-squared39   o3-Pro vs o4-mini                            0.1990    0.0005
## X-squared40   o3-Pro vs Gemini-2.5                         0.1767    0.0082
## X-squared41   o3-Pro vs Gemini-2.0-Flash                   0.3137    0.0000
## X-squared42   o3-Pro vs Gemini-2.0-Flash-GPT-Image         0.3241    0.0001
## X-squared43   o3-Pro vs All-OpenAI                         0.1735    0.0005
## X-squared44   o3-Pro vs Other-OpenAI                       0.2400    0.0000
## X-squared45   o3-Pro vs All-Gemini                         0.2610    0.0000
## X-squared46   GPT-4.1 vs GPT-4.1-GPT-Image                 0.1334    0.0482
## X-squared56   GPT-4.1-GPT-Image vs o4-mini                -0.1381    0.0172
## X-squared57   GPT-4.1-GPT-Image vs Gemini-2.5             -0.1604    0.0168
## X-squared60   GPT-4.1-GPT-Image vs All-OpenAI             -0.1636    0.0011
## X-squared76   Gemini-2.5 vs Gemini-2.0-Flash               0.1370    0.0439
## X-squared82   Gemini-2.0-Flash vs All-OpenAI              -0.1401    0.0053
## X-squared85   Gemini-2.0-Flash-GPT-Image vs All-OpenAI    -0.1506    0.0335
## X-squared88   All-OpenAI vs Other-OpenAI                   0.0665    0.0134
## X-squared89   All-OpenAI vs All-Gemini                     0.0874    0.0105
```

```r
cat("\nSignificant Comparisons in 48 Novel Tasks:\n")
```

```
##
## Significant Comparisons in 48 Novel Tasks:
```

```r
novel_sig <- novel_48_results[novel_48_results$significant, c("comparison", "diff", "p_value")]
if (nrow(novel_sig) > 0) {
  print(kable(novel_sig, format = "simple", digits = 4))
} else {
  cat("  None\n")
}
```

```
##
##
##               comparison                                            diff    p_value
## ------------  ------------------------------------------------  --------  --------
## X-squared     Humans vs o3                                       -0.0733    0.0023
## X-squared2    Humans vs o3-Pro                                   -0.1103    0.0000
## X-squared3    Humans vs GPT-4.1                                   0.1134    0.0000
## X-squared4    Humans vs GPT-4.1-GPT-Image                         0.1201    0.0000
## X-squared5    Humans vs ChatGPT-4o                                0.0998    0.0000
## X-squared7    Humans vs Gemini-2.5                                0.0719    0.0028
## X-squared8    Humans vs Gemini-2.0-Flash                          0.1232    0.0000
## X-squared9    Humans vs Gemini-2.0-Flash-GPT-Image                0.2154    0.0000
## X-squared10   Humans vs All-OpenAI                                0.0223    0.0363
## X-squared11   Humans vs Other-OpenAI                              0.0730    0.0000
## X-squared12   Humans vs All-Gemini                                0.1211    0.0000
## X-squared15   o3 vs GPT-4.1                                       0.1867    0.0000
## X-squared16   o3 vs GPT-4.1-GPT-Image                             0.1934    0.0000
## X-squared17   o3 vs ChatGPT-4o                                    0.1730    0.0000
## X-squared18   o3 vs o4-mini                                       0.0891    0.0017
## X-squared19   o3 vs Gemini-2.5                                    0.1452    0.0000
## X-squared20   o3 vs Gemini-2.0-Flash                              0.1965    0.0000
## X-squared21   o3 vs Gemini-2.0-Flash-GPT-Image                    0.2887    0.0000
## X-squared22   o3 vs All-OpenAI                                    0.0956    0.0001
## X-squared23   o3 vs Other-OpenAI                                  0.1463    0.0000
## X-squared24   o3 vs All-Gemini                                    0.1944    0.0000
## X-squared25   o3-GPT-Image vs o3-Pro                             -0.0821    0.0415
## X-squared26   o3-GPT-Image vs GPT-4.1                             0.1417    0.0004
## X-squared27   o3-GPT-Image vs GPT-4.1-GPT-Image                   0.1484    0.0002
## X-squared28   o3-GPT-Image vs ChatGPT-4o                          0.1280    0.0015
## X-squared30   o3-GPT-Image vs Gemini-2.5                          0.1002    0.0140
## X-squared31   o3-GPT-Image vs Gemini-2.0-Flash                    0.1515    0.0002
## X-squared32   o3-GPT-Image vs Gemini-2.0-Flash-GPT-Image          0.2437    0.0000
## X-squared34   o3-GPT-Image vs Other-OpenAI                        0.1013    0.0033
## X-squared35   o3-GPT-Image vs All-Gemini                          0.1494    0.0000
## X-squared36   o3-Pro vs GPT-4.1                                   0.2237    0.0000
## X-squared37   o3-Pro vs GPT-4.1-GPT-Image                         0.2305    0.0000
## X-squared38   o3-Pro vs ChatGPT-4o                                0.2101    0.0000
## X-squared39   o3-Pro vs o4-mini                                   0.1262    0.0000
## X-squared40   o3-Pro vs Gemini-2.5                                0.1822    0.0000
## X-squared41   o3-Pro vs Gemini-2.0-Flash                          0.2335    0.0000
## X-squared42   o3-Pro vs Gemini-2.0-Flash-GPT-Image                0.3257    0.0000
## X-squared43   o3-Pro vs All-OpenAI                                0.1326    0.0000
## X-squared44   o3-Pro vs Other-OpenAI                              0.1833    0.0000
## X-squared45   o3-Pro vs All-Gemini                                0.2315    0.0000
## X-squared48   GPT-4.1 vs o4-mini                                 -0.0975    0.0006
## X-squared51   GPT-4.1 vs Gemini-2.0-Flash-GPT-Image               0.1020    0.0096
## X-squared52   GPT-4.1 vs All-OpenAI                              -0.0911    0.0002
## X-squared56   GPT-4.1-GPT-Image vs o4-mini                       -0.1042    0.0002
## X-squared59   GPT-4.1-GPT-Image vs Gemini-2.0-Flash-GPT-Image     0.0953    0.0155
## X-squared60   GPT-4.1-GPT-Image vs All-OpenAI                    -0.0978    0.0001
## X-squared63   ChatGPT-4o vs o4-mini                              -0.0839    0.0032
## X-squared66   ChatGPT-4o vs Gemini-2.0-Flash-GPT-Image            0.1156    0.0034
## X-squared67   ChatGPT-4o vs All-OpenAI                           -0.0775    0.0017
## X-squared71   o4-mini vs Gemini-2.0-Flash                         0.1073    0.0001
## X-squared72   o4-mini vs Gemini-2.0-Flash-GPT-Image               0.1995    0.0000
## X-squared74   o4-mini vs Other-OpenAI                             0.0571    0.0030
## X-squared75   o4-mini vs All-Gemini                               0.1052    0.0000
## X-squared77   Gemini-2.5 vs Gemini-2.0-Flash-GPT-Image            0.1435    0.0003
## X-squared78   Gemini-2.5 vs All-OpenAI                           -0.0496    0.0463
## X-squared81   Gemini-2.0-Flash vs Gemini-2.0-Flash-GPT-Image      0.0922    0.0191
## X-squared82   Gemini-2.0-Flash vs All-OpenAI                     -0.1009    0.0000
## X-squared83   Gemini-2.0-Flash vs Other-OpenAI                   -0.0502    0.0484
## X-squared85   Gemini-2.0-Flash-GPT-Image vs All-OpenAI           -0.1931    0.0000
## X-squared86   Gemini-2.0-Flash-GPT-Image vs Other-OpenAI         -0.1424    0.0000
## X-squared87   Gemini-2.0-Flash-GPT-Image vs All-Gemini           -0.0943    0.0075
## X-squared88   All-OpenAI vs Other-OpenAI                          0.0507    0.0001
## X-squared89   All-OpenAI vs All-Gemini                            0.0988    0.0000
## X-squared90   Other-OpenAI vs All-Gemini                          0.0481    0.0066
```

Collapsed Analysis - Finke + 48 Novel Tasks Combined

```r
# Test all combinations for collapsed data
collapsed_results <- test_all_combinations(collapsed_data, "Collapsed (Finke + 48 Novel)")

# Display results
cat("All Pairwise Comparisons for Collapsed Data (Finke + 48 Novel Tasks):\n")
```

```
## All Pairwise Comparisons for Collapsed Data (Finke + 48 Novel Tasks):
```

```r
cat(paste(rep("=", 80), collapse = ""), "\n")
```

```
## ================================================================================
```

```r
for (i in 1:nrow(collapsed_results)) {
  cat("\n", collapsed_results$comparison[i], "\n")
  cat(paste(rep("-", 40), collapse = ""), "\n")
  cat("Proportions: ", round(collapsed_results$prop1[i], 3), " vs ",
      round(collapsed_results$prop2[i], 3), "\n")
  cat("Difference: ", round(collapsed_results$diff[i], 3), "\n")
  cat("Chi-squared: ", round(collapsed_results$chi_squared[i], 3), "\n")
  cat("Degrees of freedom: ", round(collapsed_results$df[i], 3), "\n")
  cat("P-value: ", format(collapsed_results$p_value[i], scientific = FALSE, digits = 4), "\n")
  cat("95% CI: [", round(collapsed_results$ci_lower[i], 3), ", ",
      round(collapsed_results$ci_upper[i], 3), "]\n")
  cat("Significant: ", ifelse(collapsed_results$significant[i], "YES (p < 0.05)", "NO"), "\n")
}
```

```
##
##  Humans vs o3
## ----------------------------------------
## Proportions:  0.539  vs  0.602
## Difference:  -0.062
## Chi-squared:  8.464
## Degrees of freedom:  1
## P-value:  0.003621
## 95% CI: [ -0.104 ,  -0.021 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o3-GPT-Image
## ----------------------------------------
## Proportions:  0.539  vs  0.541
## Difference:  -0.002
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.061 ,  0.058 ]
## Significant:  NO
##
##  Humans vs o3-Pro
## ----------------------------------------
## Proportions:  0.539  vs  0.637
## Difference:  -0.097
## Chi-squared:  20.825
## Degrees of freedom:  1
## P-value:  0.000005032
## 95% CI: [ -0.138 ,  -0.056 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-4.1
## ----------------------------------------
## Proportions:  0.539  vs  0.417
## Difference:  0.122
## Chi-squared:  32.943
## Degrees of freedom:  1
## P-value:  0.00000000949
## 95% CI: [ 0.081 ,  0.164 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.539  vs  0.385
## Difference:  0.154
## Chi-squared:  52.577
## Degrees of freedom:  1
## P-value:  0.0000000000004136
## 95% CI: [ 0.113 ,  0.196 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.539  vs  0.416
## Difference:  0.123
## Chi-squared:  33.258
## Degrees of freedom:  1
## P-value:  0.000000008073
## 95% CI: [ 0.081 ,  0.165 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o4-mini
## ----------------------------------------
## Proportions:  0.539  vs  0.496
## Difference:  0.043
## Chi-squared:  7.684
## Degrees of freedom:  1
## P-value:  0.005571
## 95% CI: [ 0.013 ,  0.074 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.539  vs  0.455
## Difference:  0.084
## Chi-squared:  15.346
## Degrees of freedom:  1
## P-value:  0.00008949
## 95% CI: [ 0.042 ,  0.126 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.539  vs  0.387
## Difference:  0.152
## Chi-squared:  51.069
## Degrees of freedom:  1
## P-value:  0.0000000000008918
## 95% CI: [ 0.111 ,  0.194 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.539  vs  0.311
## Difference:  0.228
## Chi-squared:  59.354
## Degrees of freedom:  1
## P-value:  0.00000000000001317
## 95% CI: [ 0.173 ,  0.283 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs All-OpenAI
## ----------------------------------------
## Proportions:  0.539  vs  0.496
## Difference:  0.044
## Chi-squared:  21.177
## Degrees of freedom:  1
## P-value:  0.000004187
## 95% CI: [ 0.025 ,  0.062 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.539  vs  0.442
## Difference:  0.097
## Chi-squared:  80.909
## Degrees of freedom:  1
## P-value:  0.0000000000000000002364
## 95% CI: [ 0.076 ,  0.119 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs All-Gemini
## ----------------------------------------
## Proportions:  0.539  vs  0.399
## Difference:  0.14
## Chi-squared:  97.601
## Degrees of freedom:  1
## P-value:  0.00000000000000000000005117
## 95% CI: [ 0.112 ,  0.168 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o3-GPT-Image
## ----------------------------------------
## Proportions:  0.602  vs  0.541
## Difference:  0.061
## Chi-squared:  2.797
## Degrees of freedom:  1
## P-value:  0.09442
## 95% CI: [ -0.01 ,  0.132 ]
## Significant:  NO
##
##  o3 vs o3-Pro
## ----------------------------------------
## Proportions:  0.602  vs  0.637
## Difference:  -0.035
## Chi-squared:  1.406
## Degrees of freedom:  1
## P-value:  0.2358
## 95% CI: [ -0.091 ,  0.022 ]
## Significant:  NO
##
##  o3 vs GPT-4.1
## ----------------------------------------
## Proportions:  0.602  vs  0.417
## Difference:  0.185
## Chi-squared:  40.269
## Degrees of freedom:  1
## P-value:  0.0000000002213
## 95% CI: [ 0.128 ,  0.242 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.602  vs  0.385
## Difference:  0.217
## Chi-squared:  55.585
## Degrees of freedom:  1
## P-value:  0.00000000000008952
## 95% CI: [ 0.16 ,  0.274 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.602  vs  0.416
## Difference:  0.185
## Chi-squared:  40.523
## Degrees of freedom:  1
## P-value:  0.0000000001943
## 95% CI: [ 0.128 ,  0.243 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o4-mini
## ----------------------------------------
## Proportions:  0.602  vs  0.496
## Difference:  0.106
## Chi-squared:  17.579
## Degrees of freedom:  1
## P-value:  0.00002755
## 95% CI: [ 0.056 ,  0.155 ]
## Significant:  YES (p < 0.05)
```

```
##
##  o3 vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.602  vs  0.455
## Difference:  0.146
## Chi-squared:  25.152
## Degrees of freedom:  1
## P-value:  0.0000005298
## 95% CI: [ 0.089 ,  0.204 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.602  vs  0.387
## Difference:  0.215
## Chi-squared:  54.44
## Degrees of freedom:  1
## P-value:  0.0000000000001603
## 95% CI: [ 0.158 ,  0.272 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.602  vs  0.311
## Difference:  0.291
## Chi-squared:  66.363
## Degrees of freedom:  1
## P-value:  0.000000000000000375
## 95% CI: [ 0.223 ,  0.358 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs All-OpenAI
## ----------------------------------------
## Proportions:  0.602  vs  0.496
## Difference:  0.106
## Chi-squared:  23.339
## Degrees of freedom:  1
## P-value:  0.000001358
## 95% CI: [ 0.063 ,  0.149 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.602  vs  0.442
## Difference:  0.16
## Chi-squared:  50.613
## Degrees of freedom:  1
## P-value:  0.000000000001125
## 95% CI: [ 0.116 ,  0.204 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs All-Gemini
## ----------------------------------------
## Proportions:  0.602  vs  0.399
## Difference:  0.202
## Chi-squared:  69.965
## Degrees of freedom:  1
## P-value:  0.00000000000000006037
## 95% CI: [ 0.155 ,  0.25 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs o3-Pro
## ----------------------------------------
## Proportions:  0.541  vs  0.637
## Difference:  -0.096
## Chi-squared:  7.276
## Degrees of freedom:  1
## P-value:  0.006989
## 95% CI: [ -0.167 ,  -0.025 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs GPT-4.1
## ----------------------------------------
## Proportions:  0.541  vs  0.417
## Difference:  0.124
## Chi-squared:  11.888
## Degrees of freedom:  1
## P-value:  0.0005648
## 95% CI: [ 0.053 ,  0.195 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.541  vs  0.385
## Difference:  0.156
## Chi-squared:  19.162
## Degrees of freedom:  1
## P-value:  0.00001201
## 95% CI: [ 0.085 ,  0.227 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.541  vs  0.416
## Difference:  0.125
## Chi-squared:  12.004
## Degrees of freedom:  1
## P-value:  0.000531
## 95% CI: [ 0.053 ,  0.196 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs o4-mini
## ----------------------------------------
## Proportions:  0.541  vs  0.496
## Difference:  0.045
## Chi-squared:  1.77
## Degrees of freedom:  1
## P-value:  0.1834
## 95% CI: [ -0.02 ,  0.11 ]
## Significant:  NO
##
##  o3-GPT-Image vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.541  vs  0.455
## Difference:  0.085
## Chi-squared:  5.498
## Degrees of freedom:  1
## P-value:  0.01903
## 95% CI: [ 0.014 ,  0.157 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.541  vs  0.387
## Difference:  0.154
## Chi-squared:  18.597
## Degrees of freedom:  1
## P-value:  0.00001615
## 95% CI: [ 0.083 ,  0.225 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.541  vs  0.311
## Difference:  0.23
## Chi-squared:  31.423
## Degrees of freedom:  1
## P-value:  0.00000002076
## 95% CI: [ 0.149 ,  0.31 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs All-OpenAI
## ----------------------------------------
## Proportions:  0.541  vs  0.496
## Difference:  0.045
## Chi-squared:  2.108
## Degrees of freedom:  1
## P-value:  0.1465
## 95% CI: [ -0.015 ,  0.105 ]
## Significant:  NO
##
##  o3-GPT-Image vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.541  vs  0.442
## Difference:  0.099
## Chi-squared:  10.383
## Degrees of freedom:  1
## P-value:  0.001272
## 95% CI: [ 0.038 ,  0.16 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs All-Gemini
## ----------------------------------------
## Proportions:  0.541  vs  0.399
## Difference:  0.142
## Chi-squared:  19.963
## Degrees of freedom:  1
## P-value:  0.000007895
## 95% CI: [ 0.078 ,  0.205 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-4.1
## ----------------------------------------
## Proportions:  0.637  vs  0.417
## Difference:  0.22
## Chi-squared:  57.225
## Degrees of freedom:  1
## P-value:  0.00000000000003887
## 95% CI: [ 0.163 ,  0.277 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.637  vs  0.385
## Difference:  0.252
## Chi-squared:  75.101
## Degrees of freedom:  1
## P-value:  0.000000000000000004473
## 95% CI: [ 0.195 ,  0.308 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.637  vs  0.416
## Difference:  0.22
## Chi-squared:  57.526
## Degrees of freedom:  1
## P-value:  0.00000000000003335
## 95% CI: [ 0.164 ,  0.277 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o4-mini
## ----------------------------------------
## Proportions:  0.637  vs  0.496
## Difference:  0.141
## Chi-squared:  31.377
## Degrees of freedom:  1
## P-value:  0.00000002125
## 95% CI: [ 0.092 ,  0.19 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.637  vs  0.455
## Difference:  0.181
## Chi-squared:  38.972
## Degrees of freedom:  1
## P-value:  0.0000000004299
## 95% CI: [ 0.124 ,  0.238 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.637  vs  0.387
## Difference:  0.25
## Chi-squared:  73.78
## Degrees of freedom:  1
## P-value:  0.000000000000000008734
## 95% CI: [ 0.193 ,  0.306 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.637  vs  0.311
## Difference:  0.325
## Chi-squared:  83.682
## Degrees of freedom:  1
## P-value:  0.0000000000000000000581
## 95% CI: [ 0.258 ,  0.393 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs All-OpenAI
## ----------------------------------------
## Proportions:  0.637  vs  0.496
## Difference:  0.141
## Chi-squared:  41.458
## Degrees of freedom:  1
## P-value:  0.0000000001205
## 95% CI: [ 0.099 ,  0.183 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.637  vs  0.442
## Difference:  0.195
## Chi-squared:  75.217
## Degrees of freedom:  1
## P-value:  0.00000000000000000004217
## 95% CI: [ 0.151 ,  0.238 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs All-Gemini
## ----------------------------------------
## Proportions:  0.637  vs  0.399
## Difference:  0.237
## Chi-squared:  96.046
## Degrees of freedom:  1
## P-value:  0.0000000000000000000001123
## 95% CI: [ 0.19 ,  0.284 ]
## Significant:  YES (p < 0.05)
```

```
##
##  GPT-4.1 vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.417  vs  0.385
## Difference:  0.032
## Chi-squared:  1.153
## Degrees of freedom:  1
## P-value:  0.2829
## 95% CI: [ -0.025 ,  0.089 ]
## Significant:  NO
##
##  GPT-4.1 vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.417  vs  0.416
## Difference:  0.001
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.056 ,  0.057 ]
## Significant:  NO
##
##  GPT-4.1 vs o4-mini
## ----------------------------------------
## Proportions:  0.417  vs  0.496
## Difference:  -0.079
## Chi-squared:  9.697
## Degrees of freedom:  1
## P-value:  0.001846
## 95% CI: [ -0.129 ,  -0.029 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.417  vs  0.455
## Difference:  -0.039
## Chi-squared:  1.665
## Degrees of freedom:  1
## P-value:  0.1969
## 95% CI: [ -0.096 ,  0.019 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.417  vs  0.387
## Difference:  0.03
## Chi-squared:  0.99
## Degrees of freedom:  1
## P-value:  0.3198
## 95% CI: [ -0.027 ,  0.087 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.417  vs  0.311
## Difference:  0.106
## Chi-squared:  9.023
## Degrees of freedom:  1
## P-value:  0.002666
## 95% CI: [ 0.038 ,  0.174 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs All-OpenAI
## ----------------------------------------
## Proportions:  0.417  vs  0.496
## Difference:  -0.079
## Chi-squared:  12.882
## Degrees of freedom:  1
## P-value:  0.0003318
## 95% CI: [ -0.122 ,  -0.036 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.417  vs  0.442
## Difference:  -0.025
## Chi-squared:  1.176
## Degrees of freedom:  1
## P-value:  0.2782
## 95% CI: [ -0.069 ,  0.019 ]
## Significant:  NO
##
##  GPT-4.1 vs All-Gemini
## ----------------------------------------
## Proportions:  0.417  vs  0.399
## Difference:  0.018
## Chi-squared:  0.482
## Degrees of freedom:  1
## P-value:  0.4876
## 95% CI: [ -0.03 ,  0.065 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.385  vs  0.416
## Difference:  -0.031
## Chi-squared:  1.11
## Degrees of freedom:  1
## P-value:  0.2921
## 95% CI: [ -0.089 ,  0.026 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs o4-mini
## ----------------------------------------
## Proportions:  0.385  vs  0.496
## Difference:  -0.111
## Chi-squared:  19.405
## Degrees of freedom:  1
## P-value:  0.00001058
## 95% CI: [ -0.16 ,  -0.062 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.385  vs  0.455
## Difference:  -0.071
## Chi-squared:  5.862
## Degrees of freedom:  1
## P-value:  0.01547
## 95% CI: [ -0.128 ,  -0.013 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.385  vs  0.387
## Difference:  -0.002
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  0.9842
## 95% CI: [ -0.059 ,  0.055 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.385  vs  0.311
## Difference:  0.074
## Chi-squared:  4.391
## Degrees of freedom:  1
## P-value:  0.03612
## 95% CI: [ 0.006 ,  0.141 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs All-OpenAI
## ----------------------------------------
## Proportions:  0.385  vs  0.496
## Difference:  -0.111
## Chi-squared:  25.66
## Degrees of freedom:  1
## P-value:  0.0000004071
## 95% CI: [ -0.153 ,  -0.068 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.385  vs  0.442
## Difference:  -0.057
## Chi-squared:  6.413
## Degrees of freedom:  1
## P-value:  0.01133
## 95% CI: [ -0.101 ,  -0.013 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs All-Gemini
## ----------------------------------------
## Proportions:  0.385  vs  0.399
## Difference:  -0.014
## Chi-squared:  0.315
## Degrees of freedom:  1
## P-value:  0.5744
## 95% CI: [ -0.062 ,  0.033 ]
## Significant:  NO
##
##  ChatGPT-4o vs o4-mini
## ----------------------------------------
## Proportions:  0.416  vs  0.496
## Difference:  -0.08
## Chi-squared:  9.843
## Degrees of freedom:  1
## P-value:  0.001705
## 95% CI: [ -0.129 ,  -0.03 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.416  vs  0.455
## Difference:  -0.039
## Chi-squared:  1.718
## Degrees of freedom:  1
## P-value:  0.1899
## 95% CI: [ -0.097 ,  0.019 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.416  vs  0.387
## Difference:  0.029
## Chi-squared:  0.949
## Degrees of freedom:  1
## P-value:  0.3298
## 95% CI: [ -0.028 ,  0.086 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.416  vs  0.311
## Difference:  0.105
## Chi-squared:  8.926
## Degrees of freedom:  1
## P-value:  0.002812
## 95% CI: [ 0.037 ,  0.173 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs All-OpenAI
## ----------------------------------------
## Proportions:  0.416  vs  0.496
## Difference:  -0.079
## Chi-squared:  13.074
## Degrees of freedom:  1
## P-value:  0.0002994
## 95% CI: [ -0.122 ,  -0.036 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.416  vs  0.442
## Difference:  -0.026
## Chi-squared:  1.233
## Degrees of freedom:  1
## P-value:  0.2668
## 95% CI: [ -0.07 ,  0.019 ]
## Significant:  NO
##
##  ChatGPT-4o vs All-Gemini
## ----------------------------------------
## Proportions:  0.416  vs  0.399
## Difference:  0.017
## Chi-squared:  0.449
## Degrees of freedom:  1
## P-value:  0.503
## 95% CI: [ -0.031 ,  0.065 ]
## Significant:  NO
##
##  o4-mini vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.496  vs  0.455
## Difference:  0.04
## Chi-squared:  2.448
## Degrees of freedom:  1
## P-value:  0.1177
## 95% CI: [ -0.01 ,  0.09 ]
## Significant:  NO
##
##  o4-mini vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.496  vs  0.387
## Difference:  0.109
## Chi-squared:  18.622
## Degrees of freedom:  1
## P-value:  0.00001593
## 95% CI: [ 0.059 ,  0.158 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.496  vs  0.311
## Difference:  0.185
## Chi-squared:  32.211
## Degrees of freedom:  1
## P-value:  0.00000001383
## 95% CI: [ 0.123 ,  0.246 ]
## Significant:  YES (p < 0.05)
```

```
##
##  o4-mini vs All-OpenAI
## ----------------------------------------
## Proportions:  0.496  vs  0.496
## Difference:  0
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.032 ,  0.032 ]
## Significant:  NO
##
##  o4-mini vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.496  vs  0.442
## Difference:  0.054
## Chi-squared:  9.817
## Degrees of freedom:  1
## P-value:  0.001729
## 95% CI: [ 0.02 ,  0.088 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs All-Gemini
## ----------------------------------------
## Proportions:  0.496  vs  0.399
## Difference:  0.097
## Chi-squared:  24.819
## Degrees of freedom:  1
## P-value:  0.0000006298
## 95% CI: [ 0.058 ,  0.135 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.455  vs  0.387
## Difference:  0.068
## Chi-squared:  5.487
## Degrees of freedom:  1
## P-value:  0.01916
## 95% CI: [ 0.011 ,  0.126 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5 vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.455  vs  0.311
## Difference:  0.144
## Chi-squared:  16.657
## Degrees of freedom:  1
## P-value:  0.00004478
## 95% CI: [ 0.076 ,  0.213 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5 vs All-OpenAI
## ----------------------------------------
## Proportions:  0.455  vs  0.496
## Difference:  -0.04
## Chi-squared:  3.28
## Degrees of freedom:  1
## P-value:  0.07013
## 95% CI: [ -0.084 ,  0.003 ]
## Significant:  NO
##
##  Gemini-2.5 vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.455  vs  0.442
## Difference:  0.014
## Chi-squared:  0.319
## Degrees of freedom:  1
## P-value:  0.572
## 95% CI: [ -0.031 ,  0.058 ]
## Significant:  NO
##
##  Gemini-2.5 vs All-Gemini
## ----------------------------------------
## Proportions:  0.455  vs  0.399
## Difference:  0.056
## Chi-squared:  5.352
## Degrees of freedom:  1
## P-value:  0.02069
## 95% CI: [ 0.008 ,  0.104 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.387  vs  0.311
## Difference:  0.076
## Chi-squared:  4.662
## Degrees of freedom:  1
## P-value:  0.03085
## 95% CI: [ 0.008 ,  0.144 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs All-OpenAI
## ----------------------------------------
## Proportions:  0.387  vs  0.496
## Difference:  -0.109
## Chi-squared:  24.633
## Degrees of freedom:  1
## P-value:  0.0000006935
## 95% CI: [ -0.151 ,  -0.066 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.387  vs  0.442
## Difference:  -0.055
## Chi-squared:  5.913
## Degrees of freedom:  1
## P-value:  0.01503
## 95% CI: [ -0.099 ,  -0.011 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs All-Gemini
## ----------------------------------------
## Proportions:  0.387  vs  0.399
## Difference:  -0.012
## Chi-squared:  0.218
## Degrees of freedom:  1
## P-value:  0.6403
## 95% CI: [ -0.06 ,  0.035 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-GPT-Image vs All-OpenAI
## ----------------------------------------
## Proportions:  0.311  vs  0.496
## Difference:  -0.185
## Chi-squared:  37.636
## Degrees of freedom:  1
## P-value:  0.0000000008525
## 95% CI: [ -0.241 ,  -0.128 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-GPT-Image vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.311  vs  0.442
## Difference:  -0.131
## Chi-squared:  18.49
## Degrees of freedom:  1
## P-value:  0.00001708
## 95% CI: [ -0.188 ,  -0.074 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-GPT-Image vs All-Gemini
## ----------------------------------------
## Proportions:  0.311  vs  0.399
## Difference:  -0.088
## Chi-squared:  7.824
## Degrees of freedom:  1
## P-value:  0.005156
## 95% CI: [ -0.148 ,  -0.028 ]
## Significant:  YES (p < 0.05)
##
##  All-OpenAI vs Other-OpenAI
## ----------------------------------------
## Proportions:  0.496  vs  0.442
## Difference:  0.054
## Chi-squared:  20.721
## Degrees of freedom:  1
## P-value:  0.000005312
## 95% CI: [ 0.031 ,  0.077 ]
## Significant:  YES (p < 0.05)
##
##  All-OpenAI vs All-Gemini
## ----------------------------------------
## Proportions:  0.496  vs  0.399
## Difference:  0.097
## Chi-squared:  41.68
## Degrees of freedom:  1
## P-value:  0.0000000001075
## 95% CI: [ 0.067 ,  0.126 ]
## Significant:  YES (p < 0.05)
##
##  Other-OpenAI vs All-Gemini
## ----------------------------------------
## Proportions:  0.442  vs  0.399
## Difference:  0.043
## Chi-squared:  7.268
## Degrees of freedom:  1
## P-value:  0.007021
## 95% CI: [ 0.012 ,  0.074 ]
## Significant:  YES (p < 0.05)
```

```
# Summary table
collapsed_summary <- collapsed_results %>%
  select(comparison, diff, chi_squared, p_value, significant) %>%
  mutate(diff = round(diff, 3),
         p_value = round(p_value, 4))

cat("\n\nSummary Table - Collapsed Data:\n")
```

```
##
##
## Summary Table - Collapsed Data:
```

```
print(kable(collapsed_summary, format = "simple"))
```

```
##
##
##               comparison                                           diff   chi_squared   p_value  sign
## ------------  ------------------------------------------------  -------  ------------  --------  ----
## X-squared     Humans vs o3                                       -0.062     8.4644972    0.0036  TRUE
## X-squared1    Humans vs o3-GPT-Image                             -0.002     0.0000000    1.0000  FALS
## X-squared2    Humans vs o3-Pro                                   -0.097    20.8250308    0.0000  TRUE
## X-squared3    Humans vs GPT-4.1                                   0.122    32.9430741    0.0000  TRUE
## X-squared4    Humans vs GPT-4.1-GPT-Image                         0.154    52.5773703    0.0000  TRUE
## X-squared5    Humans vs ChatGPT-4o                                0.123    33.2575489    0.0000  TRUE
## X-squared6    Humans vs o4-mini                                   0.043     7.6841989    0.0056  TRUE
## X-squared7    Humans vs Gemini-2.5                                0.084    15.3464968    0.0001  TRUE
## X-squared8    Humans vs Gemini-2.0-Flash                          0.152    51.0687955    0.0000  TRUE
## X-squared9    Humans vs Gemini-2.0-Flash-GPT-Image                0.228    59.3542594    0.0000  TRUE
## X-squared10   Humans vs All-OpenAI                                0.044    21.1771411    0.0000  TRUE
## X-squared11   Humans vs Other-OpenAI                              0.097    80.9088649    0.0000  TRUE
## X-squared12   Humans vs All-Gemini                                0.140    97.6014738    0.0000  TRUE
## X-squared13   o3 vs o3-GPT-Image                                  0.061     2.7974269    0.0944  FALS
## X-squared14   o3 vs o3-Pro                                       -0.035     1.4058191    0.2358  FALS
## X-squared15   o3 vs GPT-4.1                                       0.185    40.2685644    0.0000  TRUE
## X-squared16   o3 vs GPT-4.1-GPT-Image                             0.217    55.5846976    0.0000  TRUE
## X-squared17   o3 vs ChatGPT-4o                                    0.185    40.5228719    0.0000  TRUE
## X-squared18   o3 vs o4-mini                                       0.106    17.5794784    0.0000   TRU
## X-squared19   o3 vs Gemini-2.5                                    0.146    25.1520790    0.0000   TRU
## X-squared20   o3 vs Gemini-2.0-Flash                              0.215    54.4396550    0.0000   TRU
## X-squared21   o3 vs Gemini-2.0-Flash-GPT-Image                    0.291    66.3631960    0.0000   TRU
## X-squared22   o3 vs All-OpenAI                                    0.106    23.3388724    0.0000   TRU
## X-squared23   o3 vs Other-OpenAI                                  0.160    50.6133660    0.0000   TRU
## X-squared24   o3 vs All-Gemini                                    0.202    69.9649540    0.0000   TRU
## X-squared25   o3-GPT-Image vs o3-Pro                             -0.096     7.2757393    0.0070   TRU
## X-squared26   o3-GPT-Image vs GPT-4.1                             0.124    11.8884237    0.0006   TRU
## X-squared27   o3-GPT-Image vs GPT-4.1-GPT-Image                   0.156    19.1618578    0.0000   TRU
## X-squared28   o3-GPT-Image vs ChatGPT-4o                          0.125    12.0036649    0.0005   TRU
## X-squared29   o3-GPT-Image vs o4-mini                             0.045     1.7701281    0.1834   FALS
## X-squared30   o3-GPT-Image vs Gemini-2.5                          0.085     5.4984705    0.0190   TRU
## X-squared31   o3-GPT-Image vs Gemini-2.0-Flash                    0.154    18.5971539    0.0000   TRU
## X-squared32   o3-GPT-Image vs Gemini-2.0-Flash-GPT-Image          0.230    31.4225600    0.0000   TRU
## X-squared33   o3-GPT-Image vs All-OpenAI                          0.045     2.1081864    0.1465   FALS
## X-squared34   o3-GPT-Image vs Other-OpenAI                        0.099    10.3828764    0.0013   TRU
## X-squared35   o3-GPT-Image vs All-Gemini                          0.142    19.9630295    0.0000   TRU
## X-squared36   o3-Pro vs GPT-4.1                                   0.220    57.2251855    0.0000   TRU
## X-squared37   o3-Pro vs GPT-4.1-GPT-Image                         0.252    75.1009116    0.0000   TRU
## X-squared38   o3-Pro vs ChatGPT-4o                                0.220    57.5261185    0.0000   TRU
## X-squared39   o3-Pro vs o4-mini                                   0.141    31.3768910    0.0000   TRU
## X-squared40   o3-Pro vs Gemini-2.5                                0.181    38.9720240    0.0000   TRU
## X-squared41   o3-Pro vs Gemini-2.0-Flash                          0.250    73.7796744    0.0000   TRU
## X-squared42   o3-Pro vs Gemini-2.0-Flash-GPT-Image                0.325    83.6822547    0.0000   TRU
## X-squared43   o3-Pro vs All-OpenAI                                0.141    41.4576323    0.0000   TRU
## X-squared44   o3-Pro vs Other-OpenAI                              0.195    75.2172642    0.0000   TRU
## X-squared45   o3-Pro vs All-Gemini                                0.237    96.0455852    0.0000   TRU
## X-squared46   GPT-4.1 vs GPT-4.1-GPT-Image                        0.032     1.1529782    0.2829   FALS
## X-squared47   GPT-4.1 vs ChatGPT-4o                               0.001     0.0000000    1.0000   FALS
## X-squared48   GPT-4.1 vs o4-mini                                 -0.079     9.6969036    0.0018   TRU
## X-squared49   GPT-4.1 vs Gemini-2.5                              -0.039     1.6653880    0.1969   FALS
## X-squared50   GPT-4.1 vs Gemini-2.0-Flash                         0.030     0.9895740    0.3198   FALS
## X-squared51   GPT-4.1 vs Gemini-2.0-Flash-GPT-Image               0.106     9.0232302    0.0027   TRU
## X-squared52   GPT-4.1 vs All-OpenAI                              -0.079    12.8816785    0.0003   TRU
## X-squared53   GPT-4.1 vs Other-OpenAI                            -0.025     1.1756602    0.2782   FALS
## X-squared54   GPT-4.1 vs All-Gemini                               0.018     0.4817829    0.4876   FALS
## X-squared55   GPT-4.1-GPT-Image vs ChatGPT-4o                    -0.031     1.1096888    0.2921   FALS
## X-squared56   GPT-4.1-GPT-Image vs o4-mini                       -0.111    19.4046326    0.0000   TRU
## X-squared57   GPT-4.1-GPT-Image vs Gemini-2.5                    -0.071     5.8621757    0.0155   TRU
## X-squared58   GPT-4.1-GPT-Image vs Gemini-2.0-Flash              -0.002     0.0003907    0.9842   FALS
## X-squared59   GPT-4.1-GPT-Image vs Gemini-2.0-Flash-GPT-Image     0.074     4.3912311    0.0361   TRU
## X-squared60   GPT-4.1-GPT-Image vs All-OpenAI                    -0.111    25.6602039    0.0000   TRU
## X-squared61   GPT-4.1-GPT-Image vs Other-OpenAI                  -0.057     6.4126230    0.0113   TRU
## X-squared62   GPT-4.1-GPT-Image vs All-Gemini                    -0.014     0.3153145    0.5744   FALS
## X-squared63   ChatGPT-4o vs o4-mini                              -0.080     9.8425107    0.0017   TRU
## X-squared64   ChatGPT-4o vs Gemini-2.5                           -0.039     1.7182844    0.1899   FALS
## X-squared65   ChatGPT-4o vs Gemini-2.0-Flash                      0.029     0.9494968    0.3298   FALS
## X-squared66   ChatGPT-4o vs Gemini-2.0-Flash-GPT-Image            0.105     8.9256496    0.0028   TRU
## X-squared67   ChatGPT-4o vs All-OpenAI                           -0.079    13.0739364    0.0003   TRU
## X-squared68   ChatGPT-4o vs Other-OpenAI                         -0.026     1.2330308    0.2668   FALS
## X-squared69   ChatGPT-4o vs All-Gemini                            0.017     0.4485110    0.5030   FALS
## X-squared70   o4-mini vs Gemini-2.5                               0.040     2.4478088    0.1177   FALS
## X-squared71   o4-mini vs Gemini-2.0-Flash                         0.109    18.6222526    0.0000   TRU
## X-squared72   o4-mini vs Gemini-2.0-Flash-GPT-Image              0.185    32.2108381    0.0000  TRU
## X-squared73   o4-mini vs All-OpenAI                              0.000     0.0000000    1.0000  FALS
## X-squared74   o4-mini vs Other-OpenAI                            0.054     9.8171962    0.0017  TRU
## X-squared75   o4-mini vs All-Gemini                              0.097    24.8188928    0.0000  TRU
## X-squared76   Gemini-2.5 vs Gemini-2.0-Flash                     0.068     5.4866691    0.0192  TRU
## X-squared77   Gemini-2.5 vs Gemini-2.0-Flash-GPT-Image           0.144    16.6572049    0.0000  TRU
## X-squared78   Gemini-2.5 vs All-OpenAI                          -0.040     3.2800745    0.0701  FALS
## X-squared79   Gemini-2.5 vs Other-OpenAI                         0.014     0.3194290    0.5720  FALS
## X-squared80   Gemini-2.5 vs All-Gemini                           0.056     5.3524099    0.0207  TRU
## X-squared81   Gemini-2.0-Flash vs Gemini-2.0-Flash-GPT-Image     0.076     4.6615452    0.0308  TRU
## X-squared82   Gemini-2.0-Flash vs All-OpenAI                    -0.109    24.6331315    0.0000  TRU
## X-squared83   Gemini-2.0-Flash vs Other-OpenAI                  -0.055     5.9134847    0.0150  TRU
## X-squared84   Gemini-2.0-Flash vs All-Gemini                    -0.012     0.2183651    0.6403  FALS
## X-squared85   Gemini-2.0-Flash-GPT-Image vs All-OpenAI          -0.185    37.6362502    0.0000  TRU
## X-squared86   Gemini-2.0-Flash-GPT-Image vs Other-OpenAI        -0.131    18.4895974    0.0000  TRU
## X-squared87   Gemini-2.0-Flash-GPT-Image vs All-Gemini          -0.088     7.8238836    0.0052  TRU
## X-squared88   All-OpenAI vs Other-OpenAI                         0.054    20.7214921    0.0000  TRU
## X-squared89   All-OpenAI vs All-Gemini                           0.097    41.6804637    0.0000  TRU
## X-squared90   Other-OpenAI vs All-Gemini                         0.043     7.2677048    0.0070  TRU
```

```r
# Count significant differences
collapsed_sig_count <- sum(collapsed_results$significant)

cat("\n\nCollapsed Data Summary:\n")
```

```
##
##
## Collapsed Data Summary:
```

```r
cat("  Total comparisons:", nrow(collapsed_results), "\n")
```

```
##   Total comparisons: 91
```

```r
cat("  Significant differences:", collapsed_sig_count, "\n")
```

```
##   Significant differences: 68
```

```r
cat("  Percentage significant:", round(collapsed_sig_count / nrow(collapsed_results) * 100, 1), "%\n\n"
```

```
##   Percentage significant: 74.7 %
```

```r
# Show significant comparisons
cat("Significant Comparisons in Collapsed Data:\n")
```

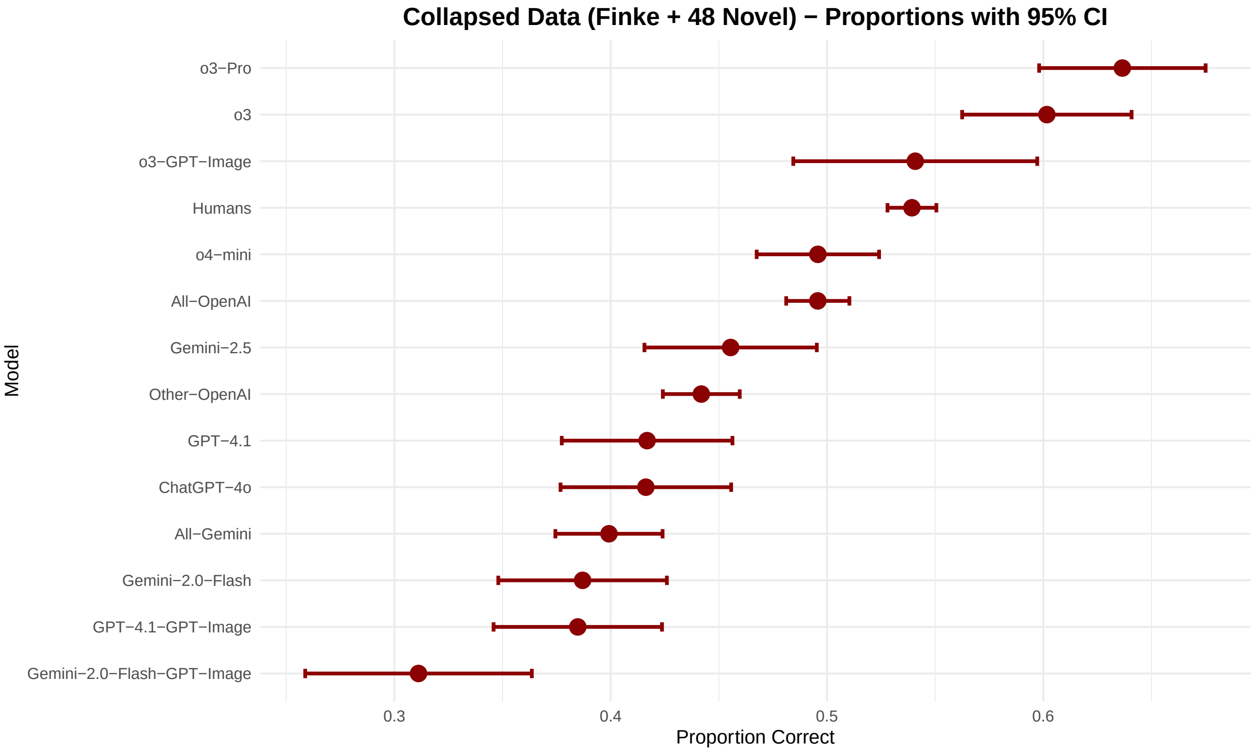
```
## Significant Comparisons in Collapsed Data:
```

```r
collapsed_sig <- collapsed_results[collapsed_results$significant, c("comparison", "diff", "p_value")]
if (nrow(collapsed_sig) > 0) {
  print(kable(collapsed_sig, format = "simple", digits = 4))
} else {
  cat("  None\n")
}
```

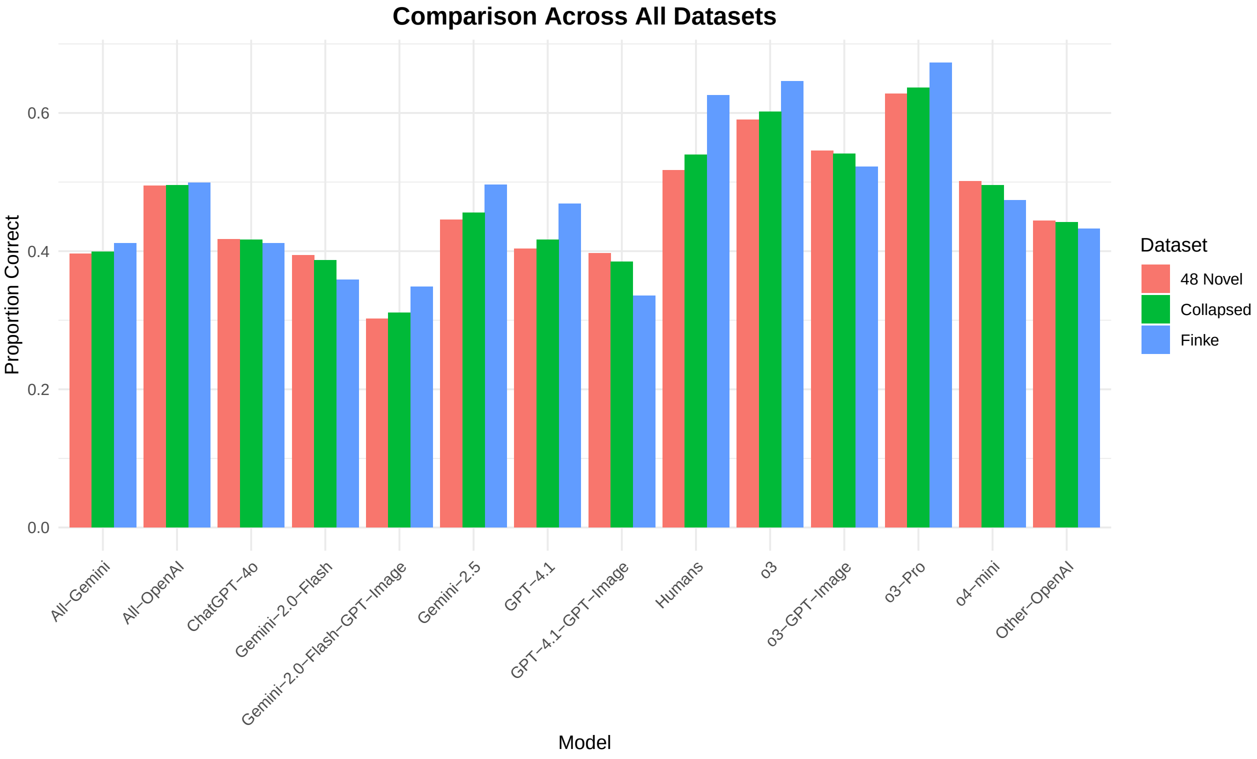
```
##
##
##                comparison                                              diff    p_value
## ------------   ------------------------------------------------   --------   --------
## X-squared      Humans vs o3                                        -0.0624     0.0036
## X-squared2     Humans vs o3-Pro                                    -0.0973     0.0000
## X-squared3    Humans vs GPT-4.1                                   0.1224    0.0000
## X-squared4    Humans vs GPT-4.1-GPT-Image                         0.1545    0.0000
## X-squared5    Humans vs ChatGPT-4o                                0.1230    0.0000
## X-squared6    Humans vs o4-mini                                   0.0435    0.0056
## X-squared7    Humans vs Gemini-2.5                                0.0838    0.0001
## X-squared8    Humans vs Gemini-2.0-Flash                          0.1523    0.0000
## X-squared9    Humans vs Gemini-2.0-Flash-GPT-Image                0.2281    0.0000
## X-squared10   Humans vs All-OpenAI                                0.0435    0.0000
## X-squared11   Humans vs Other-OpenAI                              0.0974    0.0000
## X-squared12   Humans vs All-Gemini                                0.1401    0.0000
## X-squared15   o3 vs GPT-4.1                                       0.1848    0.0000
## X-squared16   o3 vs GPT-4.1-GPT-Image                             0.2169    0.0000
## X-squared17   o3 vs ChatGPT-4o                                    0.1854    0.0000
## X-squared18   o3 vs o4-mini                                       0.1059    0.0000
## X-squared19   o3 vs Gemini-2.5                                    0.1462    0.0000
## X-squared20   o3 vs Gemini-2.0-Flash                              0.2146    0.0000
## X-squared21   o3 vs Gemini-2.0-Flash-GPT-Image                    0.2905    0.0000
## X-squared22   o3 vs All-OpenAI                                    0.1059    0.0000
## X-squared23   o3 vs Other-OpenAI                                  0.1598    0.0000
## X-squared24   o3 vs All-Gemini                                    0.2024    0.0000
## X-squared25   o3-GPT-Image vs o3-Pro                             -0.0958    0.0070
## X-squared26   o3-GPT-Image vs GPT-4.1                             0.1240    0.0006
## X-squared27   o3-GPT-Image vs GPT-4.1-GPT-Image                   0.1560    0.0000
## X-squared28   o3-GPT-Image vs ChatGPT-4o                          0.1246    0.0005
## X-squared30   o3-GPT-Image vs Gemini-2.5                          0.0854    0.0190
## X-squared31   o3-GPT-Image vs Gemini-2.0-Flash                    0.1538    0.0000
## X-squared32   o3-GPT-Image vs Gemini-2.0-Flash-GPT-Image          0.2297    0.0000
## X-squared34   o3-GPT-Image vs Other-OpenAI                        0.0989    0.0013
## X-squared35   o3-GPT-Image vs All-Gemini                          0.1416    0.0000
## X-squared36   o3-Pro vs GPT-4.1                                   0.2197    0.0000
## X-squared37   o3-Pro vs GPT-4.1-GPT-Image                         0.2518    0.0000
## X-squared38   o3-Pro vs ChatGPT-4o                                0.2203    0.0000
## X-squared39   o3-Pro vs o4-mini                                   0.1408    0.0000
## X-squared40   o3-Pro vs Gemini-2.5                                0.1811    0.0000
## X-squared41   o3-Pro vs Gemini-2.0-Flash                          0.2496    0.0000
## X-squared42   o3-Pro vs Gemini-2.0-Flash-GPT-Image                0.3254    0.0000
## X-squared43   o3-Pro vs All-OpenAI                                0.1408    0.0000
## X-squared44   o3-Pro vs Other-OpenAI                              0.1947    0.0000
## X-squared45   o3-Pro vs All-Gemini                                0.2374    0.0000
## X-squared48   GPT-4.1 vs o4-mini                                 -0.0790    0.0018
## X-squared51   GPT-4.1 vs Gemini-2.0-Flash-GPT-Image               0.1057    0.0027
## X-squared52   GPT-4.1 vs All-OpenAI                              -0.0789    0.0003
## X-squared56   GPT-4.1-GPT-Image vs o4-mini                       -0.1110    0.0000
## X-squared57   GPT-4.1-GPT-Image vs Gemini-2.5                    -0.0707    0.0155
## X-squared59   GPT-4.1-GPT-Image vs Gemini-2.0-Flash-GPT-Image     0.0736    0.0361
## X-squared60   GPT-4.1-GPT-Image vs All-OpenAI                    -0.1110    0.0000
## X-squared61   GPT-4.1-GPT-Image vs Other-OpenAI                  -0.0571    0.0113
## X-squared63   ChatGPT-4o vs o4-mini                              -0.0795    0.0017
## X-squared66   ChatGPT-4o vs Gemini-2.0-Flash-GPT-Image            0.1051    0.0028
## X-squared67   ChatGPT-4o vs All-OpenAI                           -0.0795    0.0003
## X-squared71   o4-mini vs Gemini-2.0-Flash                         0.1088    0.0000
## X-squared72   o4-mini vs Gemini-2.0-Flash-GPT-Image               0.1846    0.0000
## X-squared74   o4-mini vs Other-OpenAI                             0.0539    0.0017
## X-squared75   o4-mini vs All-Gemini                               0.0966    0.0000
## X-squared76   Gemini-2.5 vs Gemini-2.0-Flash                      0.0684    0.0192
## X-squared77   Gemini-2.5 vs Gemini-2.0-Flash-GPT-Image            0.1443    0.0000
## X-squared80   Gemini-2.5 vs All-Gemini                            0.0562    0.0207
## X-squared81   Gemini-2.0-Flash vs Gemini-2.0-Flash-GPT-Image      0.0759    0.0308
## X-squared82   Gemini-2.0-Flash vs All-OpenAI                     -0.1087    0.0000
## X-squared83   Gemini-2.0-Flash vs Other-OpenAI                   -0.0549    0.0150
## X-squared85   Gemini-2.0-Flash-GPT-Image vs All-OpenAI           -0.1846    0.0000
## X-squared86   Gemini-2.0-Flash-GPT-Image vs Other-OpenAI         -0.1307    0.0000
## X-squared87   Gemini-2.0-Flash-GPT-Image vs All-Gemini           -0.0881    0.0052
## X-squared88   All-OpenAI vs Other-OpenAI                          0.0539    0.0000
## X-squared89   All-OpenAI vs All-Gemini                            0.0965    0.0000
## X-squared90   Other-OpenAI vs All-Gemini                          0.0427    0.0070
```

Visualization of Collapsed Data

```r
# Plot proportions with confidence intervals for collapsed data
collapsed_plot <- ggplot(collapsed_data, aes(x = reorder(model, proportion), y = proportion)) +
  geom_point(size = 4, color = "darkred") +
  geom_errorbar(aes(ymin = proportion - 1.96 * sqrt(proportion * (1 - proportion) / total),
                    ymax = proportion + 1.96 * sqrt(proportion * (1 - proportion) / total)),
                width = 0.2, size = 1, color = "darkred") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Collapsed Data (Finke + 48 Novel) - Proportions with 95% CI",
       x = "Model",
       y = "Proportion Correct") +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))

print(collapsed_plot)
```

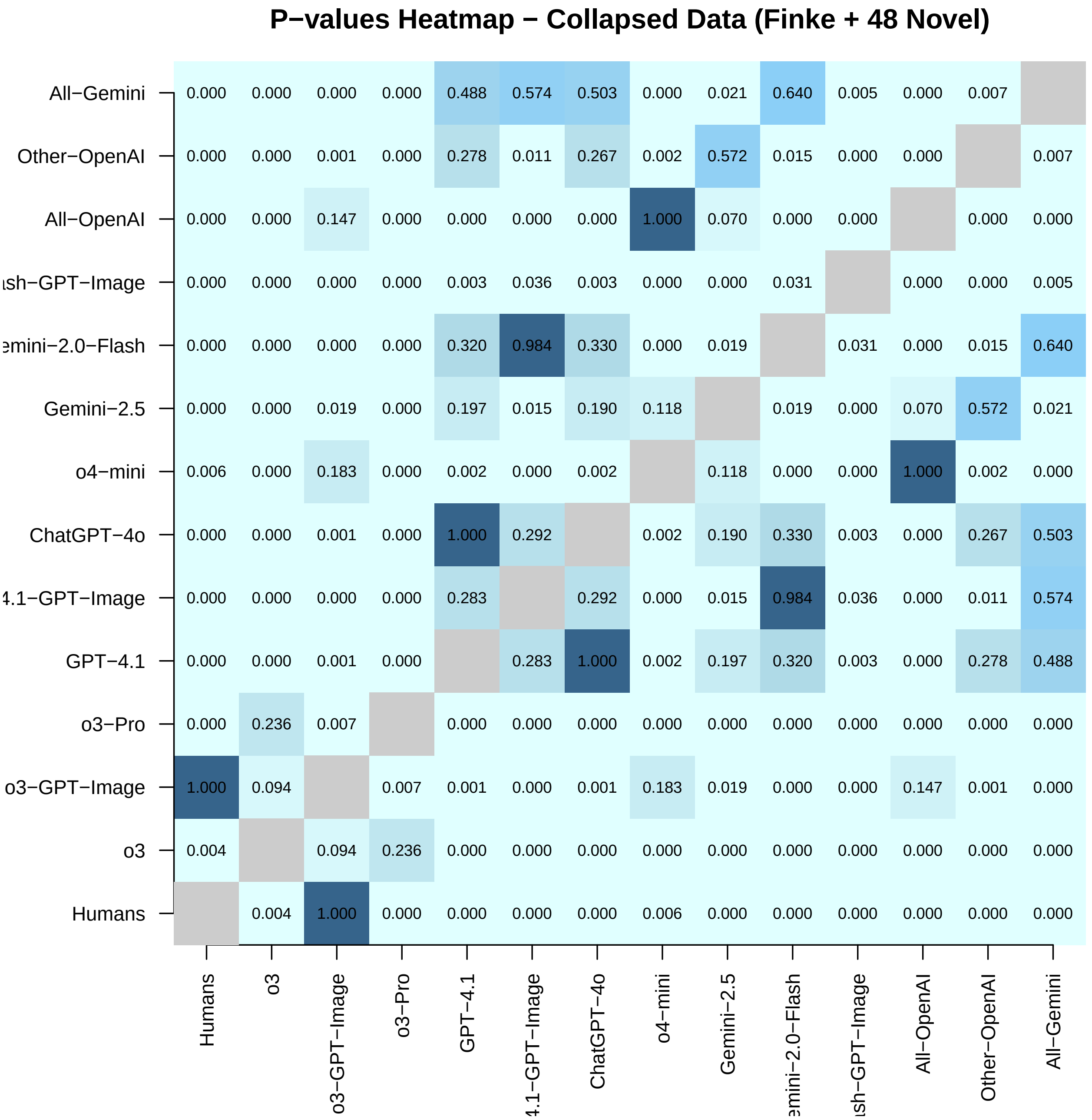


```
# Create a comparison plot showing all three datasets

comparison_data <- bind_rows(
  finke_data %>% mutate(dataset = "Finke"),
  novel_data %>% mutate(dataset = "48 Novel"),
  collapsed_data %>% mutate(dataset = "Collapsed")
)

comparison_plot <- ggplot(comparison_data, aes(x = model, y = proportion, fill = dataset)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  labs(title = "Comparison Across All Datasets",
       x = "Model",
       y = "Proportion Correct",
       fill = "Dataset") +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
        axis.text.x = element_text(angle = 45, hjust = 1))

print(comparison_plot)
```



Heatmap for Collapsed Data

```
# Create matrix of p-values for collapsed data
collapsed_models <- collapsed_data$model
collapsed_pval_matrix <- matrix(NA, nrow = length(collapsed_models), ncol = length(collapsed_models))
rownames(collapsed_pval_matrix) <- collapsed_models
colnames(collapsed_pval_matrix) <- collapsed_models

for (i in 1:nrow(collapsed_results)) {
  row_idx <- which(collapsed_models == collapsed_results$model1[i])
  col_idx <- which(collapsed_models == collapsed_results$model2[i])
  collapsed_pval_matrix[row_idx, col_idx] <- collapsed_results$p_value[i]
  collapsed_pval_matrix[col_idx, row_idx] <- collapsed_results$p_value[i]
}

# Set diagonal to NA
diag(collapsed_pval_matrix) <- NA

# Set margins for better label display
par(mar = c(6, 6, 3, 2))

# Plot heatmap with same color palette
image(collapsed_pval_matrix, axes = FALSE, col = col_palette,
      main = "P-values Heatmap - Collapsed Data (Finke + 48 Novel)")
axis(1, at = seq(0, 1, length.out = length(collapsed_models)), labels = collapsed_models,
     las = 2, cex.axis = 0.8)  # las=2 makes labels perpendicular, cex.axis makes them smaller
axis(2, at = seq(0, 1, length.out = length(collapsed_models)), labels = collapsed_models,
     las = 2, cex.axis = 0.8)

# Add gray color for diagonal
for (i in 1:length(collapsed_models)) {
  x_pos <- (i - 1) / (length(collapsed_models) - 1)
  y_pos <- (i - 1) / (length(collapsed_models) - 1)
  rect(x_pos - 0.5 / (length(collapsed_models) - 1), y_pos - 0.5 / (length(collapsed_models) - 1),
       x_pos + 0.5 / (length(collapsed_models) - 1), y_pos + 0.5 / (length(collapsed_models) - 1),
       col = "gray80", border = NA)
}

# Add p-values to the plot
for (i in 1:nrow(collapsed_pval_matrix)) {
  for (j in 1:ncol(collapsed_pval_matrix)) {
    if (!is.na(collapsed_pval_matrix[i, j])) {
      x_pos <- (j - 1) / (ncol(collapsed_pval_matrix) - 1)
      y_pos <- (i - 1) / (nrow(collapsed_pval_matrix) - 1)
      text(x_pos, y_pos, sprintf("%.3f", collapsed_pval_matrix[i, j]), cex = 0.7)
    }
  }
}
```



P−values Heatmap − Collapsed Data (Finke + 48 Novel)

	Humans	o3	o3−GPT−Image	o3−Pro	GPT−4.1	4.1−GPT−Image	ChatGPT−4o	o4−mini	Gemini−2.5	emini−2.0−Flash	sh−GPT−Image	All−OpenAI	Other−OpenAI	All−Gemini
All−Gemini	0.000	0.000	0.000	0.000	0.488	0.574	0.503	0.000	0.021	0.640	0.005	0.000	0.007	
Other−OpenAI	0.000	0.000	0.001	0.000	0.278	0.011	0.267	0.002	0.572	0.015	0.000	0.000		0.007
All−OpenAI	0.000	0.000	0.147	0.000	0.000	0.000	0.000	1.000	0.070	0.000	0.000		0.000	0.000
sh−GPT−Image	0.000	0.000	0.000	0.000	0.003	0.036	0.003	0.000	0.000	0.031		0.000	0.000	0.005
emini−2.0−Flash	0.000	0.000	0.000	0.000	0.320	0.984	0.330	0.000	0.019		0.031	0.000	0.015	0.640
Gemini−2.5	0.000	0.000	0.019	0.000	0.197	0.015	0.190	0.118		0.019	0.000	0.070	0.572	0.021
o4−mini	0.006	0.000	0.183	0.000	0.002	0.000	0.002		0.118	0.000	0.000	1.000	0.002	0.000
ChatGPT−4o	0.000	0.000	0.001	0.000	1.000	0.292		0.002	0.190	0.330	0.003	0.000	0.267	0.503
4.1−GPT−Image	0.000	0.000	0.000	0.000	0.283		0.292	0.000	0.015	0.984	0.036	0.000	0.011	0.574
GPT−4.1	0.000	0.000	0.001	0.000		0.283	1.000	0.002	0.197	0.320	0.003	0.000	0.278	0.488
o3−Pro	0.000	0.236	0.007		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
o3−GPT−Image	1.000	0.094		0.007	0.001	0.000	0.001	0.183	0.019	0.000	0.000	0.147	0.001	0.000
o3	0.004		0.094	0.236	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Humans		0.004	1.000	0.000	0.000	0.000	0.000	0.006	0.000	0.000	0.000	0.000	0.000	0.000

Export Results to CSV

```r
# Combine all results
all_results <- rbind(finke_results, novel_48_results)

# Export to CSV
write.csv(all_results, "statistical_results/proportion_test_results.csv", row.names = FALSE)
cat("\nResults exported to 'proportion_test_results.csv'\n")
```

```
##
## Results exported to 'proportion_test_results.csv'
```

```r
# Create a more detailed summary for export
detailed_summary <- all_results %>%
  mutate(
    prop1_percent = paste0(round(prop1 * 100, 1), "%"),
    prop2_percent = paste0(round(prop2 * 100, 1), "%"),
    diff_percent = paste0(round(diff * 100, 1), "%"),
    ci_95 = paste0("[", round(ci_lower, 3), ", ", round(ci_upper, 3), "]"),
    interpretation = case_when(
      p_value < 0.001 ~ "Highly significant (p < 0.001)",
      p_value < 0.01 ~ "Very significant (p < 0.01)",
      p_value < 0.05 ~ "Significant (p < 0.05)",
      p_value < 0.10 ~ "Marginally significant (p < 0.10)",
      TRUE ~ "Not significant"
    )
  ) %>%
  select(task, comparison, prop1_percent, prop2_percent, diff_percent,
         chi_squared, p_value, ci_95, interpretation)

# Export detailed summary
write.csv(detailed_summary, "statistical_results/proportion_test_detailed_summary.csv", row.names = FALS
cat("Detailed summary exported to 'proportion_test_detailed_summary.csv'\n")
```

```
## Detailed summary exported to 'proportion_test_detailed_summary.csv'
```