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

Anonymous Authors

September 23, 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 in Large Language Models.*

Please use the provided Conda environment .yml file to set up an appropriate R environment to run this R Markdown file.

```
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")

llm_data_sc_mc <- read.csv("output_csvs/single_vs_multiple_context_results.csv")
```

```
# Data
## Finke et al. Tasks - for reasoning models, only the high reasoning conditions
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$Model == "OpenAI: o3 - Single Context - High Reasoning
  llm_data_finke[llm_data_finke$Model == "OpenAI: o3 - Single Context - High Reasoning (2025-07-21)", "
  llm_data_finke[llm_data_finke$Model == "OpenAI: o3 - Multiple Context - High Reasoning (2025-09-15)",
o3_finke_max_score <- (12 + 12 + 12) * 5

o3_images_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI: o3 w/ GPT-image-1 - Multiple C
  llm_data_finke[llm_data_finke$Model == "OpenAI: o3 w/ GPT-image-1 - Multiple Context - High Reasoning
  llm_data_finke[llm_data_finke$Model == "OpenAI: o3 w/ GPT-image-1 - Multiple Context - High Reasoning
  llm_data_finke[llm_data_finke$Model == "OpenAI: o3 w/ GPT-image-1 - Multiple Context - High Reasoning
o3_images_finke_max_score <- (12 + 12 + 12 + 12) * 5

o3_pro_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI: o3 Pro - Multiple Context - High
  llm_data_finke[llm_data_finke$Model == "OpenAI: o3 Pro - Multiple Context - High Reasoning (2025-07-2
  llm_data_finke[llm_data_finke$Model == "OpenAI: o3 Pro - Multiple Context - High Reasoning (2025-09-1
o3_pro_finke_max_score <- (12 + 12 + 12) * 5

o4_mini_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI: o4-mini - Multiple Context - Hig
  llm_data_finke[llm_data_finke$Model == "OpenAI: o4-mini - Single Context - High Reasoning (2025-07-21
o4_mini_finke_max_score <- (12 + 12) * 5
```

```
chatgpt_4o_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI: ChatGPT-4o - Multiple Context
  llm_data_finke[llm_data_finke$Model == "OpenAI: ChatGPT-4o - Single Context (2025-07-25)", "overall_sc
chatgpt_4o_finke_max_score <- (12 + 12) * 5

gpt4_1_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI: GPT 4.1 - Multiple Context (2025-0
  llm_data_finke[llm_data_finke$Model == "OpenAI: GPT 4.1 - Single Context (2025-07-21)", "overall_score
gpt4_1_finke_max_score <- (12 + 12) * 5

gpt4_1_images_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI: GPT 4.1 w/ GPT-image-1 - Mu
  llm_data_finke[llm_data_finke$Model == "OpenAI: GPT 4.1 w/ GPT-Image-1 - Single Context (2025-07-21)"
gpt4_1_images_finke_max_score <- (12 + 12) * 5

gpt5_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI: GPT 5 - Multiple Context - High Rea
  llm_data_finke[llm_data_finke$Model == "OpenAI: GPT 5 - Multiple Context - High Reasoning (2025-09-15)
gpt5_finke_max_score <- (12 + 12) * 5

gemini2_5_finke_score <- llm_data_finke[llm_data_finke$Model == "DeepMind: Gemini 2.5 Pro - Multiple Co
  llm_data_finke[llm_data_finke$Model == "DeepMind: Gemini 2.5 Pro - Single Context - Dynamic Thinking
gemini2_5_finke_max_score <- (12 + 12) * 5

gemini2_0_flash_finke_score <- llm_data_finke[llm_data_finke$Model == "DeepMind: Gemini 2.0 Flash - Mult
  llm_data_finke[llm_data_finke$Model == "DeepMind: Gemini 2.0 Flash - Single Context (2025-07-21)", "ov
gemini2_0_flash_finke_max_score <- (12 + 12) * 5

gemini2_0_flash_images_finke_score <- llm_data_finke[llm_data_finke$Model == "DeepMind: Gemini 2.0 Flash
gemini2_0_flash_images_finke_max_score <- (12) * 5

opus4_1_finke_score <- llm_data_finke[llm_data_finke$Model == "Anthropic: Claude Opus 4.1 - Multiple Co
opus4_1_finke_max_score <- (12) * 5

sonnet4_finke_score <- llm_data_finke[llm_data_finke$Model == "Anthropic: Claude Sonnet 4 - Multiple Co
  llm_data_finke[llm_data_finke$Model == "Anthropic: Claude Sonnet 4 - Single Context - Extended Thinki
sonnet4_finke_max_score <- (12 + 12) * 5
```

```
## Finke Tasks - Minimal, Low, Medium Reasoning Models
medium_gpt5_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI: GPT 5 - Multiple Context - Me
medium_gpt5_finke_max_score <- (12) * 5

low_gpt5_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI: GPT 5 - Multiple Context - Low R
low_gpt5_finke_max_score <- (12) * 5

minimal_gpt5_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI: GPT 5 - Multiple Context - M
minimal_gpt5_finke_max_score <- (12) * 5

medium_o3_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI: o3 - Multiple Context - Medium
medium_o3_finke_max_score <- (12) * 5

low_o3_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI: o3 - Multiple Context - Low Reason
low_o3_finke_max_score <- (12) * 5

medium_o3_images_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI: o3 w/ GPT-image-1 - Mult
medium_o3_images_finke_max_score <- (12) * 5

medium_o4_mini_finke_score <- llm_data_finke[llm_data_finke$Model == "OpenAI: o4-mini - Multiple Contex
```

```
  llm_data_finke[llm_data_finke$Model == "OpenAI: o4-mini - Single Context - Medium Reasoning (2025-07-
medium_o4_mini_finke_max_score <- (12 + 12) * 5
```

```
## 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$Model == "OpenAI: o3 - Single Context - High Reasoning
  llm_data_novel[llm_data_novel$Model == "OpenAI: o3 - Single Context - High Reasoning (2025-07-21)", "
  llm_data_novel[llm_data_novel$Model == "OpenAI: o3 - Multiple Context - High Reasoning (2025-09-15)",
o3_novel_max_score <- (48 + 48 + 48) * 5

o3_images_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI: o3 w/ GPT-image-1 - Multiple C
  llm_data_novel[llm_data_novel$Model == "OpenAI: o3 w/ GPT-image-1 - Multiple Context - High Reasoning
  llm_data_novel[llm_data_novel$Model == "OpenAI: o3 w/ GPT-image-1 - Multiple Context - High Reasoning
  llm_data_novel[llm_data_novel$Model == "OpenAI: o3 w/ GPT-image-1 - Multiple Context - High Reasoning
o3_images_novel_max_score <- (48 + 48 + 48 + 48) * 5

o3_pro_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI: o3 Pro - Multiple Context - High
  llm_data_novel[llm_data_novel$Model == "OpenAI: o3 Pro - Multiple Context - High Reasoning (2025-07-2
  llm_data_novel[llm_data_novel$Model == "OpenAI: o3 Pro - Multiple Context - High Reasoning (2025-09-1
o3_pro_novel_max_score <- (48 + 48 + 48) * 5

o4_mini_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI: o4-mini - Multiple Context - Hig
  llm_data_novel[llm_data_novel$Model == "OpenAI: o4-mini - Single Context - High Reasoning (2025-07-21
o4_mini_novel_max_score <- (48 + 48) * 5

chatgpt_4o_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI: ChatGPT-4o - Multiple Context
  llm_data_novel[llm_data_novel$Model == "OpenAI: ChatGPT-4o - Single Context (2025-07-25)", "overall_s
chatgpt_4o_novel_max_score <- (48 + 48) * 5

gpt4_1_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI: GPT 4.1 - Multiple Context (2025-
  llm_data_novel[llm_data_novel$Model == "OpenAI: GPT 4.1 - Single Context (2025-07-21)", "overall_scor
gpt4_1_novel_max_score <- (48 + 48) * 5

gpt4_1_images_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI: GPT 4.1 w/ GPT-image-1 - M
  llm_data_novel[llm_data_novel$Model == "OpenAI: GPT 4.1 w/ GPT-Image-1 - Single Context (2025-07-21)"
gpt4_1_images_novel_max_score <- (48 + 48) * 5

gpt5_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI: GPT 5 - Multiple Context - High Rea
  llm_data_novel[llm_data_novel$Model == "OpenAI: GPT 5 - Multiple Context - High Reasoning (2025-09-15
gpt5_novel_max_score <- (48 + 48) * 5

gemini2_5_novel_score <- llm_data_novel[llm_data_novel$Model == "DeepMind: Gemini 2.5 Pro - Multiple Co
  llm_data_novel[llm_data_novel$Model == "DeepMind: Gemini 2.5 Pro - Single Context - Dynamic Thinking
gemini2_5_novel_max_score <- (48 + 48) * 5

gemini2_0_flash_novel_score <- llm_data_novel[llm_data_novel$Model == "DeepMind: Gemini 2.0 Flash - Mul
  llm_data_novel[llm_data_novel$Model == "DeepMind: Gemini 2.0 Flash - Single Context (2025-07-21)", "o
gemini2_0_flash_novel_max_score <- (48 + 48) * 5

gemini2_0_flash_images_novel_score <- llm_data_novel[llm_data_novel$Model == "DeepMind: Gemini 2.0 Flas
gemini2_0_flash_images_novel_max_score <- (48) * 5
```

```r
opus4_1_novel_score <- llm_data_novel[llm_data_novel$Model == "Anthropic: Claude Opus 4.1 - Multiple Co
opus4_1_novel_max_score <- (48) * 5

sonnet4_novel_score <- llm_data_novel[llm_data_novel$Model == "Anthropic: Claude Sonnet 4 - Multiple Co
  llm_data_novel[llm_data_novel$Model == "Anthropic: Claude Sonnet 4 - Single Context - Extended Thinki
sonnet4_novel_max_score <- (48 + 48) * 5
```

```r
## Novel Tasks - Minimal, Low, Medium Reasoning Models
medium_gpt5_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI: GPT 5 - Multiple Context - M
medium_gpt5_novel_max_score <- (48) * 5

low_gpt5_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI: GPT 5 - Multiple Context - Low
low_gpt5_novel_max_score <- (48) * 5

minimal_gpt5_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI: GPT 5 - Multiple Context -
minimal_gpt5_novel_max_score <- (48) * 5

medium_o3_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI: o3 - Multiple Context - Medium
medium_o3_novel_max_score <- (48) * 5

low_o3_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI: o3 - Multiple Context - Low Reaso
low_o3_novel_max_score <- (48) * 5

medium_o3_images_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI: o3 w/ GPT-image-1 - Mul
medium_o3_images_novel_max_score <- (48) * 5

medium_o4_mini_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI: o4-mini - Multiple Contex
  llm_data_novel[llm_data_novel$Model == "OpenAI: o4-mini - Single Context - Medium Reasoning (2025-07-
medium_o4_mini_novel_max_score <- (48 + 48) * 5
```

```r
o3_collapsed_sc <- llm_data_sc_mc[llm_data_sc_mc$Model == "o3_sc", "overall_score"]
o3_collapsed_sc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "o3_sc", "n_total"]) * 5

o3_collapsed_mc <- llm_data_sc_mc[llm_data_sc_mc$Model == "o3_mc", "overall_score"]
o3_collapsed_mc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "o3_mc", "n_total"]) * 5


o3_pro_collapsed_sc <- llm_data_sc_mc[llm_data_sc_mc$Model == "o3_pro_sc", "overall_score"]
o3_pro_collapsed_sc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "o3_pro_sc", "n_total"]) * 5

o3_pro_collapsed_mc <- llm_data_sc_mc[llm_data_sc_mc$Model == "o3_pro_mc", "overall_score"]
o3_pro_collapsed_mc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "o3_pro_mc", "n_total"]) * 5


o4_mini_collapsed_sc <- llm_data_sc_mc[llm_data_sc_mc$Model == "o4_mini_sc", "overall_score"]
o4_mini_collapsed_sc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "o4_mini_sc", "n_total"]) * 5

o4_mini_collapsed_mc <- llm_data_sc_mc[llm_data_sc_mc$Model == "o4_mini_mc", "overall_score"]
o4_mini_collapsed_mc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "o4_mini_mc", "n_total"]) * 5


sonnet_collapsed_sc <- llm_data_sc_mc[llm_data_sc_mc$Model == "sonnet_sc", "overall_score"]
sonnet_collapsed_sc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "sonnet_sc", "n_total"]) * 5
```

```
sonnet_collapsed_mc <- llm_data_sc_mc[llm_data_sc_mc$Model == "sonnet_mc", "overall_score"]
sonnet_collapsed_mc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "sonnet_mc", "n_total"]) * 5


gemini2_0_flash_sc <- llm_data_sc_mc[llm_data_sc_mc$Model == "gemini_2.0_flash_sc", "overall_score"]
gemini2_0_flash_sc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "gemini_2.0_flash_sc", "n_total"]) *

gemini2_0_flash_mc <- llm_data_sc_mc[llm_data_sc_mc$Model == "gemini_2.0_flash_mc", "overall_score"]
gemini2_0_flash_mc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "gemini_2.0_flash_mc", "n_total"]) *


gemini2_5_pro_sc <- llm_data_sc_mc[llm_data_sc_mc$Model == "gemini_2.5_pro_sc", "overall_score"]
gemini2_5_pro_sc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "gemini_2.5_pro_sc", "n_total"]) * 5

gemini2_5_pro_mc <- llm_data_sc_mc[llm_data_sc_mc$Model == "gemini_2.5_pro_mc", "overall_score"]
gemini2_5_pro_mc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "gemini_2.5_pro_mc", "n_total"]) * 5


chatgpt4o_collapsed_sc <- llm_data_sc_mc[llm_data_sc_mc$Model == "chatgpt4o_sc", "overall_score"]
chatgpt4o_collapsed_sc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "chatgpt4o_sc", "n_total"]) * 5

chatgpt4o_collapsed_mc <- llm_data_sc_mc[llm_data_sc_mc$Model == "chatgpt4o_mc", "overall_score"]
chatgpt4o_collapsed_mc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "chatgpt4o_mc", "n_total"]) * 5


gpt4_1_collapsed_sc <- llm_data_sc_mc[llm_data_sc_mc$Model == "gpt4.1_sc", "overall_score"]
gpt4_1_collapsed_sc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "gpt4.1_sc", "n_total"]) * 5

gpt4_1_collapsed_mc <- llm_data_sc_mc[llm_data_sc_mc$Model == "gpt4.1_mc", "overall_score"]
gpt4_1_collapsed_mc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "gpt4.1_mc", "n_total"]) * 5


gpt_4_1_images_collapsed_sc <- llm_data_sc_mc[llm_data_sc_mc$Model == "gpt4.1_images_sc", "overall_scor
gpt_4_1_images_collapsed_sc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "gpt4.1_images_sc", "n_total

gpt_4_1_images_collapsed_mc <- llm_data_sc_mc[llm_data_sc_mc$Model == "gpt4.1_images_mc", "overall_scor
gpt_4_1_images_collapsed_mc_max <- (llm_data_sc_mc[llm_data_sc_mc$Model == "gpt4.1_images_mc", "n_total



total_collapsed_sc <- o3_collapsed_sc +
  o3_pro_collapsed_sc +
  o4_mini_collapsed_sc +
  sonnet_collapsed_sc +
  gemini2_0_flash_sc +
  gemini2_5_pro_sc +
  chatgpt4o_collapsed_sc +
  gpt4_1_collapsed_sc +
  gpt_4_1_images_collapsed_sc
total_collapsed_sc_max <- o3_collapsed_sc_max +
  o3_pro_collapsed_sc_max +
  o4_mini_collapsed_sc_max +
  sonnet_collapsed_sc_max +
```

```r
  gemini2_0_flash_sc_max +
  gemini2_5_pro_sc_max +
  chatgpt4o_collapsed_sc_max +
  gpt4_1_collapsed_sc_max +
  gpt_4_1_images_collapsed_sc_max

total_collapsed_mc <- o3_collapsed_mc +
  o3_pro_collapsed_mc +
  o4_mini_collapsed_mc +
  sonnet_collapsed_mc +
  gemini2_0_flash_mc +
  gemini2_5_pro_mc +
  chatgpt4o_collapsed_mc +
  gpt4_1_collapsed_mc +
  gpt_4_1_images_collapsed_mc
total_collapsed_mc_max <- o3_collapsed_mc_max +
  o3_pro_collapsed_mc_max +
  o4_mini_collapsed_mc_max +
  sonnet_collapsed_mc_max +
  gemini2_0_flash_mc_max +
  gemini2_5_pro_mc_max +
  chatgpt4o_collapsed_mc_max +
  gpt4_1_collapsed_mc_max +
  gpt_4_1_images_collapsed_mc_max
```

```r
## 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

gpt5_total_score <- gpt5_finke_score + gpt5_novel_score
gpt5_total_max_score <- gpt5_finke_max_score + gpt5_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_image

opus4_1_total_score <- opus4_1_finke_score + opus4_1_novel_score
opus4_1_total_max_score <- opus4_1_finke_max_score + opus4_1_novel_max_score

sonnet4_total_score <- sonnet4_finke_score + sonnet4_novel_score
sonnet4_total_max_score <- sonnet4_finke_max_score + sonnet4_novel_max_score


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

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
```

```
## Collapsed Data - Minimal, Low, Medium Reasoning Models
medium_gpt5_total_score <- medium_gpt5_finke_score + medium_gpt5_novel_score
medium_gpt5_total_max_score <- medium_gpt5_finke_max_score + medium_gpt5_novel_max_score

low_gpt5_total_score <- low_gpt5_finke_score + low_gpt5_novel_score
low_gpt5_total_max_score <- low_gpt5_finke_max_score + low_gpt5_novel_max_score

minimal_gpt5_total_score <- minimal_gpt5_finke_score + minimal_gpt5_novel_score
minimal_gpt5_total_max_score <- minimal_gpt5_finke_max_score + minimal_gpt5_novel_max_score

medium_o3_total_score <- medium_o3_finke_score + medium_o3_novel_score
medium_o3_total_max_score <- medium_o3_finke_max_score + medium_o3_novel_max_score

low_o3_total_score <- low_o3_finke_score + low_o3_novel_score
low_o3_total_max_score <- low_o3_finke_max_score + low_o3_novel_max_score

medium_o3_images_total_score <- medium_o3_images_finke_score + medium_o3_images_novel_score
medium_o3_images_total_max_score <- medium_o3_images_finke_max_score + medium_o3_images_novel_max_score

medium_o4_mini_total_score <- medium_o4_mini_finke_score + medium_o4_mini_novel_score
medium_o4_mini_total_max_score <- medium_o4_mini_finke_max_score + medium_o4_mini_novel_max_score
```

```
# Create data frames for easier manipulation
sc_mc_data <- data.frame(
  model = c("o3-SC", "o3-MC",
            "o3-Pro-SC", "o3-Pro-MC",
            "o4-mini-SC", "o4-mini-MC",
            "Sonnet-4-SC", "Sonnet-4-MC",
            "Gemini-2.0-Flash-SC", "Gemini-2.0-Flash-MC",
            "Gemini-2.5-Pro-SC", "Gemini-2.5-Pro-MC",
            "ChatGPT-4o-SC", "ChatGPT-4o-MC",
            "GPT-4.1-SC", "GPT-4.1-MC",
            "GPT-4.1-GPT-Image-SC", "GPT-4.1-GPT-Image-MC"),
  score = c(o3_collapsed_sc, o3_collapsed_mc,
            o3_pro_collapsed_sc, o3_pro_collapsed_mc,
            o4_mini_collapsed_sc, o4_mini_collapsed_mc,
            sonnet_collapsed_sc, sonnet_collapsed_mc,
            gemini2_0_flash_sc, gemini2_0_flash_mc,
            gemini2_5_pro_sc, gemini2_5_pro_mc,
            chatgpt4o_collapsed_sc, chatgpt4o_collapsed_mc,
            gpt4_1_collapsed_sc, gpt4_1_collapsed_mc,
            gpt_4_1_images_collapsed_sc, gpt_4_1_images_collapsed_mc),
  max_score = c(o3_collapsed_sc_max, o3_collapsed_mc_max,
                o3_pro_collapsed_sc_max, o3_pro_collapsed_mc_max,
                o4_mini_collapsed_sc_max, o4_mini_collapsed_mc_max,
                sonnet_collapsed_sc_max, sonnet_collapsed_mc_max,
                gemini2_0_flash_sc_max, gemini2_0_flash_mc_max,
                gemini2_5_pro_sc_max, gemini2_5_pro_mc_max,
                chatgpt4o_collapsed_sc_max, chatgpt4o_collapsed_mc_max,
                gpt4_1_collapsed_sc_max, gpt4_1_collapsed_mc_max,
                gpt_4_1_images_collapsed_sc_max, gpt_4_1_images_collapsed_mc_max),
  color = c("#fc8d62", "#fc8d62",
            "#fc8d62", "#fc8d62",
            "#fc8d62", "#fc8d62",
            "#e78ac3", "#e78ac3",
            "#8da0cb", "#8da0cb",
            "#8da0cb", "#8da0cb",
            "#fc8d62", "#fc8d62",
            "#fc8d62", "#fc8d62",
            "#fc8d62", "#fc8d62"),

  shape = c(16, 18,
            16, 18,
            16, 18,
            16, 18,
            16, 18,
            16, 18,
            16, 18,
            16, 18,
            16, 18)
)
# Calculate proportions from correct/total
sc_mc_data$proportion <- sc_mc_data$score / sc_mc_data$max_score
```

```
# 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-Pro",
            "Gemini-2.0-Flash", "Gemini-2.0-Flash-Images",
            "Sonnet-4", "Opus-4.1", "GPT-5"),
  score = c(humans_finke_score, o3_finke_score, o3_images_finke_score,
```

```r
            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,
            sonnet4_finke_score, opus4_1_finke_score, gpt5_finke_score),
  max_score = 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,
                sonnet4_finke_max_score, opus4_1_finke_max_score, gpt5_finke_max_score),
  color = c("#66c2a5", "#fc8d62", "#fc8d62",
            "#fc8d62", "#fc8d62", "#fc8d62",
            "#fc8d62", "#fc8d62", "#8da0cb",
            "#8da0cb", "#8da0cb", "#e78ac3", "#e78ac3", "#fc8d62")

  # human #66c2a5
  # openai #fc8d62
  # gemini #8da0cb
  # claude #e78ac3
)

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

```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-Pro",
            "Gemini-2.0-Flash", "Gemini-2.0-Flash-Images",
            "Sonnet-4", "Opus-4.1", "GPT-5"),
  score = 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,
            sonnet4_novel_score, opus4_1_novel_score, gpt5_novel_score),
  max_score = 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,
                sonnet4_novel_max_score, opus4_1_novel_max_score, gpt5_novel_max_score),
  color = c("#66c2a5", "#fc8d62", "#fc8d62",
            "#fc8d62", "#fc8d62", "#fc8d62",
            "#fc8d62", "#fc8d62", "#8da0cb",
            "#8da0cb", "#8da0cb", "#e78ac3", "#e78ac3", "#fc8d62")

)

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

```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-Pro",
            "Gemini-2.0-Flash", "Gemini-2.0-Flash-Images",
            "Sonnet-4", "Opus-4.1", "GPT-5"),
  score = 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,
            sonnet4_total_score, opus4_1_total_score, gpt5_total_score),
  max_score = 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,
                sonnet4_total_max_score, opus4_1_total_max_score, gpt5_total_max_score),
  color = c("#66c2a5", "#fc8d62", "#fc8d62",
            "#fc8d62", "#fc8d62", "#fc8d62",
            "#fc8d62", "#fc8d62", "#8da0cb",
            "#8da0cb", "#8da0cb", "#e78ac3", "#e78ac3", "#fc8d62")
)

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

Set up Data for Reasoning Variations

```
# Prepare data for reasoning variations analysis
finke_reasoning_data <- data.frame(
  model = c("Humans", "o3-High", "o3-Medium",
            "o3-Low", 'GPT-5-High', 'o3-Pro',
            "GPT-5-Medium", "GPT-5-Low", "GPT-5-Minimal",
            "o4-mini-High", "o4-mini-Medium", "o3-GPT-Image-High",
            "o3-GPT-Image-Medium"),
  score = c(humans_finke_score, o3_finke_score, medium_o3_finke_score,
            low_o3_finke_score, gpt5_finke_score, o3_pro_finke_score,
            medium_gpt5_finke_score, low_gpt5_finke_score, minimal_gpt5_finke_score,
            o4_mini_finke_score, medium_o4_mini_finke_score, o3_images_finke_score,
            medium_o3_images_finke_score),
  max_score = c(humans_finke_max_score, o3_finke_max_score, medium_o3_finke_max_score,
                low_o3_finke_max_score, gpt5_finke_max_score, o3_pro_finke_max_score,
                medium_gpt5_finke_max_score, low_gpt5_finke_max_score,
                minimal_gpt5_finke_max_score, o4_mini_finke_max_score, medium_o4_mini_finke_max_score, o
                medium_o3_images_finke_max_score),
  color = c("#66c2a5", "#980043", "#dd1c77",
            "#df65b0", "#980043", "#980043",
            "#dd1c77", "#df65b0",
            "#d7b5d8", "#980043", "#dd1c77", "#980043", "#dd1c77")
)

# Calculate proportions from score/max_score
finke_reasoning_data$proportion <- finke_reasoning_data$score / finke_reasoning_data$max_score
```

```
novel_reasoning_data <- data.frame(
  model = c("Humans", "o3-High", "o3-Medium",
            "o3-Low", 'GPT-5-High', 'o3-Pro',
            "GPT-5-Medium", "GPT-5-Low", "GPT-5-Minimal",
            "o4-mini-High", "o4-mini-Medium", "o3-GPT-Image-High",
```

```r
            "o3-GPT-Image-Medium"),
  score = c(humans_novel_score, o3_novel_score, medium_o3_novel_score,
            low_o3_novel_score,
            gpt5_novel_score, medium_gpt5_novel_score, o3_pro_novel_score, low_gpt5_novel_score,
            minimal_gpt5_novel_score, o4_mini_novel_score, medium_o4_mini_novel_score,
            o3_images_novel_score, medium_o3_images_novel_score),
  max_score = c(humans_novel_max_score, o3_novel_max_score, medium_o3_novel_max_score,
                low_o3_novel_max_score,
                gpt5_novel_max_score, medium_gpt5_novel_max_score, o3_pro_novel_max_score, low_gpt5_nove
                minimal_gpt5_novel_max_score, o4_mini_novel_max_score, medium_o4_mini_novel_max_score,
                o3_images_novel_max_score, medium_o3_images_novel_max_score),
  color = c("#66c2a5", "#980043", "#dd1c77",
            "#df65b0", "#980043", "#980043",
            "#dd1c77", "#df65b0",
            "#d7b5d8", "#980043", "#dd1c77", "#980043", "#dd1c77")
)
# Calculate proportions from score/max_score
novel_reasoning_data$proportion <- novel_reasoning_data$score / novel_reasoning_data$max_score
```

```r
collapsed_reasoning_data <- data.frame(
  model = c("Humans", "o3-High", "o3-Medium",
            "o3-Low", 'GPT-5-High', 'o3-Pro',
            "GPT-5-Medium", "GPT-5-Low", "GPT-5-Minimal",
            "o4-mini-High", "o4-mini-Medium", "o3-GPT-Image-High",
            "o3-GPT-Image-Medium"),
  score = c(humans_total_score, o3_total_score, medium_o3_total_score,
            low_o3_total_score,
            gpt5_total_score, o3_pro_total_score, medium_gpt5_total_score, low_gpt5_total_score,
            minimal_gpt5_total_score, o4_mini_total_score, medium_o4_mini_total_score,
            o3_images_total_score, medium_o3_images_total_score),
  max_score = c(humans_total_max_score, o3_total_max_score, medium_o3_total_max_score,
                low_o3_total_max_score,
                gpt5_total_max_score, o3_pro_total_max_score, medium_gpt5_total_max_score, low_gpt5_tota
                minimal_gpt5_total_max_score, o4_mini_total_max_score, medium_o4_mini_total_max_score,
                o3_images_total_max_score, medium_o3_images_total_max_score),
  color = c("#66c2a5", "#980043", "#dd1c77",
            "#df65b0", "#980043", "#980043",
            "#dd1c77", "#df65b0",
            "#d7b5d8", "#980043", "#dd1c77", "#980043", "#dd1c77")
)
# Calculate proportions from score/max_score
collapsed_reasoning_data$proportion <- collapsed_reasoning_data$score / collapsed_reasoning_data$max_sco
```

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

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

```r
print(finke_data)
```

```
##                 model     score max_score   color proportion
## 1              Humans 961.09643      1525 #66c2a5  0.6302272
## 2                  o3 109.90000       180 #fc8d62  0.6105556
## 3         o3-GPT-Image 134.48333       240 #fc8d62  0.5603472
## 4              o3-Pro 138.90833       180 #fc8d62  0.7717130
## 5                   GPT-4.1  56.40714       120 #fc8d62  0.4700595
## 6         GPT-4.1-GPT-Image  41.00000       120 #fc8d62  0.3416667
## 7                ChatGPT-4o  48.98095       120 #fc8d62  0.4081746
## 8                   o4-mini  63.00833       120 #fc8d62  0.5250694
## 9            Gemini-2.5-Pro  61.12500       120 #8da0cb  0.5093750
## 10         Gemini-2.0-Flash  41.10000       120 #8da0cb  0.3425000
## 11 Gemini-2.0-Flash-Images  20.53810        60 #8da0cb  0.3423016
## 12                 Sonnet-4  54.65238       120 #e78ac3  0.4554365
## 13                 Opus-4.1  44.46667        60 #e78ac3  0.7411111
## 14                    GPT-5  91.95000       120 #fc8d62  0.7662500
```

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

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

```
print(novel_data)
```

```
##                       model      score max_score   color proportion
## 1                    Humans 3137.12024      5965 #66c2a5  0.5259212
## 2                        o3  467.49048       720 #fc8d62  0.6492923
## 3              o3-GPT-Image  529.69881       960 #fc8d62  0.5517696
## 4                    o3-Pro  460.71310       720 #fc8d62  0.6398793
## 5                   GPT-4.1  198.45476       480 #fc8d62  0.4134474
## 6         GPT-4.1-GPT-Image  188.82738       480 #fc8d62  0.3933904
## 7                ChatGPT-4o  202.76786       480 #fc8d62  0.4224330
## 8                   o4-mini  255.12262       480 #fc8d62  0.5315055
## 9            Gemini-2.5-Pro  215.94881       480 #8da0cb  0.4498934
## 10         Gemini-2.0-Flash  186.88214       480 #8da0cb  0.3893378
## 11 Gemini-2.0-Flash-Images   78.80714       240 #8da0cb  0.3283631
## 12                 Sonnet-4  195.48810       480 #e78ac3  0.4072669
## 13                 Opus-4.1  114.35238       240 #e78ac3  0.4764683
## 14                    GPT-5  309.87262       480 #fc8d62  0.6455680
```

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

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

```
print(collapsed_data)
```

```
##                       model      score max_score   color proportion
## 1                    Humans 4098.21667      7490 #66c2a5  0.5471584
## 2                        o3  577.39048       900 #fc8d62  0.6415450
## 3              o3-GPT-Image  664.18214      1200 #fc8d62  0.5534851
## 4                    o3-Pro  599.62143       900 #fc8d62  0.6662460
## 5                   GPT-4.1  254.86190       600 #fc8d62  0.4247698
## 6         GPT-4.1-GPT-Image  229.82738       600 #fc8d62  0.3830456
## 7                ChatGPT-4o  251.74881       600 #fc8d62  0.4195813
## 8                   o4-mini  318.13095       600 #fc8d62  0.5302183
## 9            Gemini-2.5-Pro  277.07381       600 #8da0cb  0.4617897
## 10         Gemini-2.0-Flash  227.98214       600 #8da0cb  0.3799702
## 11 Gemini-2.0-Flash-Images   99.34524       300 #8da0cb  0.3311508
## 12                 Sonnet-4  250.14048       600 #e78ac3  0.4169008
## 13                 Opus-4.1  158.81905       300 #e78ac3  0.5293968
## 14                    GPT-5  401.82262       600 #fc8d62  0.6697044
```

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

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

```r
cat("Exp 2: ", original_finke_exp2_correct, "/", original_finke_exp2_total, " (", round(original_finke_
```

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

```r
cat("Exp 3: ", original_finke_exp3_correct, "/", original_finke_exp3_total, " (", round(original_finke_
```

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

```r
cat("Collapsed Original Finke: ", original_finke_correct, "/", original_finke_total, " (", round(origina
```

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

```r
# Display the reasoning variation data
cat("\n\nFinke et al. Tasks - Reasoning Variations Data:\n")
```

```
##
##
## Finke et al. Tasks - Reasoning Variations Data:
```

```r
print(finke_reasoning_data)
```

```
##                  model     score max_score   color proportion
## 1               Humans 961.09643      1525 #66c2a5  0.6302272
## 2              o3-High 109.90000       180 #980043  0.6105556
## 3            o3-Medium  34.41667        60 #dd1c77  0.5736111
## 4               o3-Low  37.38333        60 #df65b0  0.6230556
## 5           GPT-5-High  91.95000       120 #980043  0.7662500
## 6               o3-Pro 138.90833       180 #980043  0.7717130
## 7         GPT-5-Medium  38.00833        60 #dd1c77  0.6334722
## 8            GPT-5-Low  33.35833        60 #df65b0  0.5559722
## 9        GPT-5-Minimal  21.93452        60 #d7b5d8  0.3655754
## 10        o4-mini-High  63.00833       120 #980043  0.5250694
## 11      o4-mini-Medium  56.02500       120 #dd1c77  0.4668750
## 12   o3-GPT-Image-High 134.48333       240 #980043  0.5603472
## 13 o3-GPT-Image-Medium  30.33810        60 #dd1c77  0.5056349
```

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

```
##
## 48 Novel Tasks - Reasoning Variations Data:
```

```r
print(novel_reasoning_data)
```

```
##                  model     score max_score   color proportion
## 1               Humans 3137.1202      5965 #66c2a5  0.5259212
## 2              o3-High  467.4905       720 #980043  0.6492923
## 3            o3-Medium  134.8440       240 #dd1c77  0.5618502
## 4               o3-Low  124.4119       240 #df65b0  0.5183829
## 5           GPT-5-High  309.8726       480 #980043  0.6455680
## 6               o3-Pro  140.3917       240 #980043  0.5849653
## 7         GPT-5-Medium  460.7131       720 #dd1c77  0.6398793
## 8            GPT-5-Low  118.2940       240 #df65b0  0.4928919
## 9        GPT-5-Minimal  100.2702       240 #d7b5d8  0.4177927
## 10        o4-mini-High  255.1226       480 #980043  0.5315055
## 11      o4-mini-Medium  237.6810       480 #dd1c77  0.4951687
## 12   o3-GPT-Image-High  529.6988       960 #980043  0.5517696
## 13 o3-GPT-Image-Medium  134.2131       240 #dd1c77  0.5592212
```

```r
cat("\nCollapsed Data (Finke + 48 Novel Tasks) - Reasoning Variations Data:\n")
```

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

```r
print(collapsed_reasoning_data)
```

```
##                  model     score max_score   color proportion
## 1               Humans 4098.2167      7490 #66c2a5  0.5471584
## 2              o3-High  577.3905       900 #980043  0.6415450
## 3            o3-Medium  169.2607       300 #dd1c77  0.5642024
## 4               o3-Low  161.7952       300 #df65b0  0.5393175
## 5           GPT-5-High  401.8226       600 #980043  0.6697044
## 6               o3-Pro  599.6214       900 #980043  0.6662460
## 7         GPT-5-Medium  178.4000       300 #dd1c77  0.5946667
## 8            GPT-5-Low  151.6524       300 #df65b0  0.5055079
## 9        GPT-5-Minimal  122.2048       300 #d7b5d8  0.4073492
## 10        o4-mini-High  318.1310       600 #980043  0.5302183
## 11      o4-mini-Medium  293.7060       600 #dd1c77  0.4895099
## 12   o3-GPT-Image-High  664.1821      1200 #980043  0.5534851
## 13 o3-GPT-Image-Medium  164.5512       300 #dd1c77  0.5485040
```

Proportion Testing Function

```r
# Function to perform proportion test and extract results
perform_prop_test <- function(model1_name, model1_correct, model1_total,
                              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$model[i], data$score[i], data$max_score[i],
        data$model[j], data$score[j], data$max_score[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: Current Human Finke vs Original Finke

```
##
##
## Comparison: Current Human Finke vs Original Finke (Collapsed Exp 2 + Exp 3)

## ======================================================================

## Current Human Finke: 961.0964/1525 (0.63)

## Original Finke: 404/720 (0.561)

## Difference:  0.069

## Chi-squared:  9.516

## P-value:  0.002037

## 95% CI: [ 0.024 ,  0.114 ]

## Significant:  YES (p < 0.05)

##
##
## Detailed Comparison: Current Humans vs Original Finke

## ----------------------------------------
## Proportions:  0.63  vs  0.561
## Difference:  0.069
## Chi-squared:  9.516
## Degrees of freedom:  1
## P-value:  0.002037
## 95% CI: [ 0.024 ,  0.114 ]
## Significant:  YES (p < 0.05)
##
##
## Summary Table - Human vs Original Finke:
##
##
## comparison                           diff   p_value  significant
## ---------------------------------  ------  --------  ------------
## Current Humans vs Original Finke    0.069     0.002  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: 3137.12/5965 (0.526)
## Original Finke: 404/720 (0.561)
## Difference:  -0.035
## Chi-squared:  3.054
## P-value:  0.08055
## 95% CI: [ -0.074 ,  0.004 ]
## Significant:  NO
##
##
## Detailed Comparison: Current Humans vs Original Finke
## ----------------------------------------
## Proportions:  0.526  vs  0.561
## Difference:  -0.035
## Chi-squared:  3.054
## Degrees of freedom:  1
## P-value:  0.08055
## 95% CI: [ -0.074 ,  0.004 ]
## Significant:  NO
```

```
##
##
## Summary Table - Human vs Original Finke:
##
##
## comparison                            diff   p_value  significant
## ---------------------------------  -------  --------  ------------
## Current Humans vs Original Finke    -0.035    0.0805  FALSE
```

Comparison: Current Humans (collapsed) vs. Original Finke

```
##
##
## Comparison: Current Human 48-Item Task vs Original Finke (Collapsed Exp 2 + Exp 3)
## ======================================================================
## Current Human Finke: 4098.217/7490 (0.547)
## Original Finke: 404/720 (0.561)
## Difference:  -0.014
## Chi-squared:  0.462
## P-value:  0.4969
## 95% CI: [ -0.053 ,  0.025 ]
## Significant:  NO
##
##
## Detailed Comparison: Current Humans vs Original Finke
## ----------------------------------------
## Proportions:  0.547  vs  0.561
## Difference:  -0.014
## Chi-squared:  0.462
## Degrees of freedom:  1
## P-value:  0.4969
## 95% CI: [ -0.053 ,  0.025 ]
## Significant:  NO
##
##
## Summary Table - Current Human (Collapsed) vs Original Finke:
##
##
## comparison                                        diff   p_value  significant
## ---------------------------------------------  -------  --------  ------------
## Current Humans (collapsed) vs Original Finke    -0.014    0.4969  FALSE
```

Single Context vs Multiple Context - All Pairwise Comparisons

```
## All Pairwise Comparisons for Single-Context vs Multiple-Context:
## ==================================================================================
##
##  o3-SC vs o3-MC
## ----------------------------------------
## Proportions:  0.636  vs  0.622
## Difference:  0.014
## Chi-squared:  0.106
## Degrees of freedom:  1
## P-value:  0.7443
## 95% CI: [ -0.056 ,  0.083 ]
## Significant:  NO
##
##  o3-SC vs o3-Pro-SC
## ----------------------------------------
## Proportions:  0.636  vs  0.667
## Difference:  -0.032
## Chi-squared:  0.524
## Degrees of freedom:  1
## P-value:  0.469
## 95% CI: [ -0.111 ,  0.048 ]
## Significant:  NO
##
##  o3-SC vs o3-Pro-MC
## ----------------------------------------
## Proportions:  0.636  vs  0.66
## Difference:  -0.024
## Chi-squared:  0.417
## Degrees of freedom:  1
## P-value:  0.5183
## 95% CI: [ -0.093 ,  0.045 ]
## Significant:  NO
##
##  o3-SC vs o4-mini-SC
## ----------------------------------------
## Proportions:  0.636  vs  0.487
## Difference:  0.149
## Chi-squared:  12.868
## Degrees of freedom:  1
## P-value:  0.0003342
## 95% CI: [ 0.067 ,  0.231 ]
## Significant:  YES (p < 0.05)
##
##  o3-SC vs o4-mini-MC
## ----------------------------------------
## Proportions:  0.636  vs  0.573
## Difference:  0.063
## Chi-squared:  2.216
## Degrees of freedom:  1
## P-value:  0.1366
## 95% CI: [ -0.019 ,  0.144 ]
## Significant:  NO
##
##  o3-SC vs Sonnet-4-SC
## ----------------------------------------
## Proportions:  0.636  vs  0.428
## Difference:  0.208
## Chi-squared:  25.307
## Degrees of freedom:  1
## P-value:  0.000000489
## 95% CI: [ 0.127 ,  0.29 ]
## Significant:  YES (p < 0.05)
##
##  o3-SC vs Sonnet-4-MC
## ----------------------------------------
## Proportions:  0.636  vs  0.406
## Difference:  0.23
## Chi-squared:  30.82
## Degrees of freedom:  1
## P-value:  0.00000002831
## 95% CI: [ 0.149 ,  0.311 ]
## Significant:  YES (p < 0.05)
##
##  o3-SC vs Gemini-2.0-Flash-SC
## ----------------------------------------
## Proportions:  0.636  vs  0.387
## Difference:  0.249
## Chi-squared:  36.123
## Degrees of freedom:  1
## P-value:  0.000000001852
## 95% CI: [ 0.168 ,  0.329 ]
## Significant:  YES (p < 0.05)
##
##  o3-SC vs Gemini-2.0-Flash-MC
## ----------------------------------------
## Proportions:  0.636  vs  0.373
## Difference:  0.263
## Chi-squared:  40.549
## Degrees of freedom:  1
## P-value:  0.0000000001917
## 95% CI: [ 0.183 ,  0.344 ]
## Significant:  YES (p < 0.05)
##
##  o3-SC vs Gemini-2.5-Pro-SC
## ----------------------------------------
## Proportions:  0.636  vs  0.467
## Difference:  0.169
## Chi-squared:  16.577
## Degrees of freedom:  1
## P-value:  0.00004672
## 95% CI: [ 0.087 ,  0.25 ]
## Significant:  YES (p < 0.05)
##
##  o3-SC vs Gemini-2.5-Pro-MC
## ----------------------------------------
## Proportions:  0.636  vs  0.456
## Difference:  0.18
## Chi-squared:  18.807
## Degrees of freedom:  1
## P-value:  0.00001446
## 95% CI: [ 0.098 ,  0.261 ]
## Significant:  YES (p < 0.05)
##
##  o3-SC vs ChatGPT-4o-SC
## ----------------------------------------
## Proportions:  0.636  vs  0.398
## Difference:  0.238
## Chi-squared:  33.06
## Degrees of freedom:  1
## P-value:  0.000000008935
## 95% CI: [ 0.157 ,  0.319 ]
## Significant:  YES (p < 0.05)
##
##  o3-SC vs ChatGPT-4o-MC
## ----------------------------------------
## Proportions:  0.636  vs  0.441
## Difference:  0.195
## Chi-squared:  22.12
## Degrees of freedom:  1
## P-value:  0.000002561
## 95% CI: [ 0.113 ,  0.276 ]
## Significant:  YES (p < 0.05)
##
##  o3-SC vs GPT-4.1-SC
## ----------------------------------------
## Proportions:  0.636  vs  0.441
## Difference:  0.195
## Chi-squared:  22.068
## Degrees of freedom:  1
## P-value:  0.000002632
## 95% CI: [ 0.113 ,  0.276 ]
## Significant:  YES (p < 0.05)
##
##  o3-SC vs GPT-4.1-MC
## ----------------------------------------
## Proportions:  0.636  vs  0.408
## Difference:  0.228
## Chi-squared:  30.286
## Degrees of freedom:  1
## P-value:  0.00000003728
## 95% CI: [ 0.147 ,  0.309 ]
## Significant:  YES (p < 0.05)
##
##  o3-SC vs GPT-4.1-GPT-Image-SC
## ----------------------------------------
## Proportions:  0.636  vs  0.386
## Difference:  0.25
## Chi-squared:  36.467
## Degrees of freedom:  1
## P-value:  0.000000001553
## 95% CI: [ 0.169 ,  0.331 ]
## Significant:  YES (p < 0.05)
##
##  o3-SC vs GPT-4.1-GPT-Image-MC
## ----------------------------------------
## Proportions:  0.636  vs  0.38
## Difference:  0.256
## Chi-squared:  38.304
## Degrees of freedom:  1
## P-value:  0.0000000006053
## 95% CI: [ 0.175 ,  0.337 ]
## Significant:  YES (p < 0.05)
##
##  o3-MC vs o3-Pro-SC
## ----------------------------------------
## Proportions:  0.622  vs  0.667
## Difference:  -0.045
## Chi-squared:  1.575
## Degrees of freedom:  1
## P-value:  0.2095
## 95% CI: [ -0.114 ,  0.023 ]
## Significant:  NO
##
##  o3-MC vs o3-Pro-MC
## ----------------------------------------
## Proportions:  0.622  vs  0.66
## Difference:  -0.038
## Chi-squared:  1.713
## Degrees of freedom:  1
## P-value:  0.1906
## 95% CI: [ -0.094 ,  0.018 ]
## Significant:  NO
##
##  o3-MC vs o4-mini-SC
## ----------------------------------------
## Proportions:  0.622  vs  0.487
## Difference:  0.135
## Chi-squared:  14.391
## Degrees of freedom:  1
## P-value:  0.0001485
## 95% CI: [ 0.064 ,  0.206 ]
## Significant:  YES (p < 0.05)
##
##  o3-MC vs o4-mini-MC
## ----------------------------------------
## Proportions:  0.622  vs  0.573
## Difference:  0.049
## Chi-squared:  1.819
## Degrees of freedom:  1
## P-value:  0.1774
## 95% CI: [ -0.021 ,  0.12 ]
## Significant:  NO
##
```

```
##  o3-MC vs Sonnet-4-SC
## ----------------------------------------
## Proportions:  0.622  vs  0.428
## Difference:  0.195
## Chi-squared:  29.926
## Degrees of freedom:  1
## P-value:  0.00000004488
## 95% CI: [ 0.124 ,  0.265 ]
## Significant:  YES (p < 0.05)
##
##  o3-MC vs Sonnet-4-MC
## ----------------------------------------
## Proportions:  0.622  vs  0.406
## Difference:  0.216
## Chi-squared:  36.878
## Degrees of freedom:  1
## P-value:  0.000000001258
## 95% CI: [ 0.146 ,  0.286 ]
## Significant:  YES (p < 0.05)
##
##  o3-MC vs Gemini-2.0-Flash-SC
## ----------------------------------------
## Proportions:  0.622  vs  0.387
## Difference:  0.235
## Chi-squared:  43.574
## Degrees of freedom:  1
## P-value:  0.00000000004083
## 95% CI: [ 0.165 ,  0.305 ]
## Significant:  YES (p < 0.05)
##
##  o3-MC vs Gemini-2.0-Flash-MC
## ----------------------------------------
## Proportions:  0.622  vs  0.373
## Difference:  0.25
## Chi-squared:  49.16
## Degrees of freedom:  1
## P-value:  0.000000000002359
## 95% CI: [ 0.18 ,  0.319 ]
## Significant:  YES (p < 0.05)
##
##  o3-MC vs Gemini-2.5-Pro-SC
## ----------------------------------------
## Proportions:  0.622  vs  0.467
## Difference:  0.155
## Chi-squared:  18.985
## Degrees of freedom:  1
## P-value:  0.00001317
## 95% CI: [ 0.084 ,  0.226 ]
## Significant:  YES (p < 0.05)
##
##  o3-MC vs Gemini-2.5-Pro-MC
## ----------------------------------------
## Proportions:  0.622  vs  0.456
## Difference:  0.166
## Chi-squared:  21.767
## Degrees of freedom:  1
## P-value:  0.000003078
## 95% CI: [ 0.095 ,  0.237 ]
## Significant:  YES (p < 0.05)
##
##  o3-MC vs ChatGPT-4o-SC
## ----------------------------------------
## Proportions:  0.622  vs  0.398
## Difference:  0.224
## Chi-squared:  39.706
## Degrees of freedom:  1
## P-value:  0.0000000002952
## 95% CI: [ 0.154 ,  0.294 ]
## Significant:  YES (p < 0.05)
##
##  o3-MC vs ChatGPT-4o-MC
## ----------------------------------------
## Proportions:  0.622  vs  0.441
## Difference:  0.181
## Chi-squared:  25.919
## Degrees of freedom:  1
## P-value:  0.000000356
## 95% CI: [ 0.11 ,  0.252 ]
## Significant:  YES (p < 0.05)
##
##  o3-MC vs GPT-4.1-SC
## ----------------------------------------
## Proportions:  0.622  vs  0.441
## Difference:  0.181
## Chi-squared:  25.854
## Degrees of freedom:  1
## P-value:  0.0000003683
## 95% CI: [ 0.11 ,  0.252 ]
## Significant:  YES (p < 0.05)
##
##  o3-MC vs GPT-4.1-MC
## ----------------------------------------
## Proportions:  0.622  vs  0.408
## Difference:  0.214
## Chi-squared:  36.205
## Degrees of freedom:  1
## P-value:  0.000000001776
## 95% CI: [ 0.144 ,  0.284 ]
## Significant:  YES (p < 0.05)
##
##  o3-MC vs GPT-4.1-GPT-Image-SC
## ----------------------------------------
## Proportions:  0.622  vs  0.386
## Difference:  0.236
## Chi-squared:  44.007
## Degrees of freedom:  1
## P-value:  0.00000000003271
## 95% CI: [ 0.166 ,  0.306 ]
## Significant:  YES (p < 0.05)
##
##  o3-MC vs GPT-4.1-GPT-Image-MC
## ----------------------------------------
## Proportions:  0.622  vs  0.38
## Difference:  0.242
## Chi-squared:  46.327
## Degrees of freedom:  1
## P-value:  0.00000000001001
## 95% CI: [ 0.173 ,  0.312 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-SC vs o3-Pro-MC
## ----------------------------------------
## Proportions:  0.667  vs  0.66
## Difference:  0.007
## Chi-squared:  0.02
## Degrees of freedom:  1
## P-value:  0.887
## 95% CI: [ -0.061 ,  0.075 ]
## Significant:  NO
##
##  o3-Pro-SC vs o4-mini-SC
## ----------------------------------------
## Proportions:  0.667  vs  0.487
## Difference:  0.18
## Chi-squared:  19.223
## Degrees of freedom:  1
## P-value:  0.00001163
## 95% CI: [ 0.099 ,  0.261 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-SC vs o4-mini-MC
## ----------------------------------------
## Proportions:  0.667  vs  0.573
## Difference:  0.094
## Chi-squared:  5.266
## Degrees of freedom:  1
## P-value:  0.02174
## 95% CI: [ 0.014 ,  0.175 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-SC vs Sonnet-4-SC
## ----------------------------------------
## Proportions:  0.667  vs  0.428
## Difference:  0.24
## Chi-squared:  33.854
## Degrees of freedom:  1
## P-value:  0.00000000594
## 95% CI: [ 0.159 ,  0.32 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-SC vs Sonnet-4-MC
## ----------------------------------------
## Proportions:  0.667  vs  0.406
## Difference:  0.261
## Chi-squared:  40.139
## Degrees of freedom:  1
## P-value:  0.0000000002366
## 95% CI: [ 0.181 ,  0.342 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-SC vs Gemini-2.0-Flash-SC
## ----------------------------------------
## Proportions:  0.667  vs  0.387
## Difference:  0.28
## Chi-squared:  46.111
## Degrees of freedom:  1
## P-value:  0.00000000001117
## 95% CI: [ 0.2 ,  0.36 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-SC vs Gemini-2.0-Flash-MC
## ----------------------------------------
## Proportions:  0.667  vs  0.373
## Difference:  0.295
## Chi-squared:  51.051
## Degrees of freedom:  1
## P-value:  0.0000000000008998
## 95% CI: [ 0.215 ,  0.375 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-SC vs Gemini-2.5-Pro-SC
## ----------------------------------------
## Proportions:  0.667  vs  0.467
## Difference:  0.2
## Chi-squared:  23.676
## Degrees of freedom:  1
## P-value:  0.00000114
## 95% CI: [ 0.119 ,  0.281 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-SC vs Gemini-2.5-Pro-MC
## ----------------------------------------
## Proportions:  0.667  vs  0.456
## Difference:  0.211
## Chi-squared:  26.31
## Degrees of freedom:  1
## P-value:  0.0000002908
## 95% CI: [ 0.13 ,  0.292 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-SC vs ChatGPT-4o-SC
## ----------------------------------------
## Proportions:  0.667  vs  0.398
## Difference:  0.269
## Chi-squared:  42.669
## Degrees of freedom:  1
## P-value:  0.00000000006482
## 95% CI: [ 0.189 ,  0.35 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-SC vs ChatGPT-4o-MC
## ----------------------------------------
## Proportions:  0.667  vs  0.441
## Difference:  0.226
## Chi-squared:  30.177
## Degrees of freedom:  1
## P-value:  0.00000003943
## 95% CI: [ 0.145 ,  0.307 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-SC vs GPT-4.1-SC
## ----------------------------------------
## Proportions:  0.667  vs  0.441
## Difference:  0.226
## Chi-squared:  30.117
## Degrees of freedom:  1
## P-value:  0.00000004068
## 95% CI: [ 0.145 ,  0.307 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-SC vs GPT-4.1-MC
## ----------------------------------------
## Proportions:  0.667  vs  0.408
## Difference:  0.259
## Chi-squared:  39.534
## Degrees of freedom:  1
## P-value:  0.0000000003224
## 95% CI: [ 0.179 ,  0.34 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-SC vs GPT-4.1-GPT-Image-SC
## ----------------------------------------
## Proportions:  0.667  vs  0.386
## Difference:  0.281
## Chi-squared:  46.496
## Degrees of freedom:  1
## P-value:  0.000000000009181
## 95% CI: [ 0.201 ,  0.361 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-SC vs GPT-4.1-GPT-Image-MC
## ----------------------------------------
## Proportions:  0.667  vs  0.38
## Difference:  0.287
## Chi-squared:  48.55
## Degrees of freedom:  1
## P-value:  0.00000000000322
## 95% CI: [ 0.208 ,  0.367 ]
## Significant:  YES (p < 0.05)
##
```

```
##  o3-Pro-MC vs o4-mini-SC
## ----------------------------------------
## Proportions:  0.66  vs  0.487
## Difference:  0.173
## Chi-squared:  24.255
## Degrees of freedom:  1
## P-value:  0.000000844
## 95% CI: [ 0.102 ,  0.244 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-MC vs o4-mini-MC
## ----------------------------------------
## Proportions:  0.66  vs  0.573
## Difference:  0.087
## Chi-squared:  6.137
## Degrees of freedom:  1
## P-value:  0.01324
## 95% CI: [ 0.017 ,  0.157 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-MC vs Sonnet-4-SC
## ----------------------------------------
## Proportions:  0.66  vs  0.428
## Difference:  0.233
## Chi-squared:  43.526
## Degrees of freedom:  1
## P-value:  0.00000000004184
## 95% CI: [ 0.162 ,  0.303 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-MC vs Sonnet-4-MC
## ----------------------------------------
## Proportions:  0.66  vs  0.406
## Difference:  0.254
## Chi-squared:  51.794
## Degrees of freedom:  1
## P-value:  0.0000000000006163
## 95% CI: [ 0.184 ,  0.324 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-MC vs Gemini-2.0-Flash-SC
## ----------------------------------------
## Proportions:  0.66  vs  0.387
## Difference:  0.273
## Chi-squared:  59.631
## Degrees of freedom:  1
## P-value:  0.00000000000001144
## 95% CI: [ 0.203 ,  0.342 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-MC vs Gemini-2.0-Flash-MC
## ----------------------------------------
## Proportions:  0.66  vs  0.373
## Difference:  0.288
## Chi-squared:  66.094
## Degrees of freedom:  1
## P-value:  0.0000000000000004299
## 95% CI: [ 0.218 ,  0.357 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-MC vs Gemini-2.5-Pro-SC
## ----------------------------------------
## Proportions:  0.66  vs  0.467
## Difference:  0.193
## Chi-squared:  30.114
## Degrees of freedom:  1
## P-value:  0.00000004073
## 95% CI: [ 0.122 ,  0.263 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-MC vs Gemini-2.5-Pro-MC
## ----------------------------------------
## Proportions:  0.66  vs  0.456
## Difference:  0.204
## Chi-squared:  33.585
## Degrees of freedom:  1
## P-value:  0.000000006822
## 95% CI: [ 0.133 ,  0.274 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-MC vs ChatGPT-4o-SC
## ----------------------------------------
## Proportions:  0.66  vs  0.398
## Difference:  0.262
## Chi-squared:  55.118
## Degrees of freedom:  1
## P-value:  0.0000000000001135
## 95% CI: [ 0.193 ,  0.332 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-MC vs ChatGPT-4o-MC
## ----------------------------------------
## Proportions:  0.66  vs  0.441
## Difference:  0.219
## Chi-squared:  38.682
## Degrees of freedom:  1
## P-value:  0.0000000004988
## 95% CI: [ 0.149 ,  0.289 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-MC vs GPT-4.1-SC
## ----------------------------------------
## Proportions:  0.66  vs  0.441
## Difference:  0.219
## Chi-squared:  38.602
## Degrees of freedom:  1
## P-value:  0.0000000005197
## 95% CI: [ 0.149 ,  0.289 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-MC vs GPT-4.1-MC
## ----------------------------------------
## Proportions:  0.66  vs  0.408
## Difference:  0.252
## Chi-squared:  51
## Degrees of freedom:  1
## P-value:  0.0000000000009238
## 95% CI: [ 0.182 ,  0.322 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-MC vs GPT-4.1-GPT-Image-SC
## ----------------------------------------
## Proportions:  0.66  vs  0.386
## Difference:  0.274
## Chi-squared:  60.135
## Degrees of freedom:  1
## P-value:  0.000000000000008855
## 95% CI: [ 0.205 ,  0.343 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro-MC vs GPT-4.1-GPT-Image-MC
## ----------------------------------------
## Proportions:  0.66  vs  0.38
## Difference:  0.28
## Chi-squared:  62.824
## Degrees of freedom:  1
## P-value:  0.00000000000000226
## 95% CI: [ 0.211 ,  0.349 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-SC vs o4-mini-MC
## ----------------------------------------
## Proportions:  0.487  vs  0.573
## Difference:  -0.086
## Chi-squared:  4.106
## Degrees of freedom:  1
## P-value:  0.04274
## 95% CI: [ -0.169 ,  -0.003 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-SC vs Sonnet-4-SC
## ----------------------------------------
## Proportions:  0.487  vs  0.428
## Difference:  0.06
## Chi-squared:  1.915
## Degrees of freedom:  1
## P-value:  0.1664
## 95% CI: [ -0.023 ,  0.143 ]
## Significant:  NO
##
##  o4-mini-SC vs Sonnet-4-MC
## ----------------------------------------
## Proportions:  0.487  vs  0.406
## Difference:  0.081
## Chi-squared:  3.671
## Degrees of freedom:  1
## P-value:  0.05537
## 95% CI: [ -0.002 ,  0.164 ]
## Significant:  NO
##
##  o4-mini-SC vs Gemini-2.0-Flash-SC
## ----------------------------------------
## Proportions:  0.487  vs  0.387
## Difference:  0.1
## Chi-squared:  5.693
## Degrees of freedom:  1
## P-value:  0.01704
## 95% CI: [ 0.018 ,  0.182 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-SC vs Gemini-2.0-Flash-MC
## ----------------------------------------
## Proportions:  0.487  vs  0.373
## Difference:  0.115
## Chi-squared:  7.58
## Degrees of freedom:  1
## P-value:  0.005902
## 95% CI: [ 0.033 ,  0.197 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-SC vs Gemini-2.5-Pro-SC
## ----------------------------------------
## Proportions:  0.487  vs  0.467
## Difference:  0.02
## Chi-squared:  0.167
## Degrees of freedom:  1
## P-value:  0.6829
## 95% CI: [ -0.063 ,  0.103 ]
## Significant:  NO
##
##  o4-mini-SC vs Gemini-2.5-Pro-MC
## ----------------------------------------
## Proportions:  0.487  vs  0.456
## Difference:  0.031
## Chi-squared:  0.459
## Degrees of freedom:  1
## P-value:  0.498
## 95% CI: [ -0.052 ,  0.114 ]
## Significant:  NO
##
##  o4-mini-SC vs ChatGPT-4o-SC
## ----------------------------------------
## Proportions:  0.487  vs  0.398
## Difference:  0.089
## Chi-squared:  4.49
## Degrees of freedom:  1
## P-value:  0.0341
## 95% CI: [ 0.007 ,  0.172 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-SC vs ChatGPT-4o-MC
## ----------------------------------------
## Proportions:  0.487  vs  0.441
## Difference:  0.046
## Chi-squared:  1.103
## Degrees of freedom:  1
## P-value:  0.2936
## 95% CI: [ -0.037 ,  0.129 ]
## Significant:  NO
##
##  o4-mini-SC vs GPT-4.1-SC
## ----------------------------------------
## Proportions:  0.487  vs  0.441
## Difference:  0.046
## Chi-squared:  1.091
## Degrees of freedom:  1
## P-value:  0.2962
## 95% CI: [ -0.037 ,  0.129 ]
## Significant:  NO
##
##  o4-mini-SC vs GPT-4.1-MC
## ----------------------------------------
## Proportions:  0.487  vs  0.408
## Difference:  0.079
## Chi-squared:  3.484
## Degrees of freedom:  1
## P-value:  0.06196
## 95% CI: [ -0.004 ,  0.162 ]
## Significant:  NO
##
##  o4-mini-SC vs GPT-4.1-GPT-Image-SC
## ----------------------------------------
## Proportions:  0.487  vs  0.386
## Difference:  0.101
## Chi-squared:  5.833
## Degrees of freedom:  1
## P-value:  0.01573
## 95% CI: [ 0.019 ,  0.183 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-SC vs GPT-4.1-GPT-Image-MC
## ----------------------------------------
## Proportions:  0.487  vs  0.38
## Difference:  0.107
## Chi-squared:  6.602
## Degrees of freedom:  1
## P-value:  0.01019
## 95% CI: [ 0.025 ,  0.189 ]
## Significant:  YES (p < 0.05)
##
```

```
##  o4-mini-MC vs Sonnet-4-SC
## ----------------------------------------
## Proportions:  0.573  vs  0.428
## Difference:  0.146
## Chi-squared:  12.132
## Degrees of freedom:  1
## P-value:  0.0004958
## 95% CI: [ 0.063 ,  0.228 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-MC vs Sonnet-4-MC
## ----------------------------------------
## Proportions:  0.573  vs  0.406
## Difference:  0.167
## Chi-squared:  16.081
## Degrees of freedom:  1
## P-value:  0.00006068
## 95% CI: [ 0.085 ,  0.249 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-MC vs Gemini-2.0-Flash-SC
## ----------------------------------------
## Proportions:  0.573  vs  0.387
## Difference:  0.186
## Chi-squared:  20.024
## Degrees of freedom:  1
## P-value:  0.000007649
## 95% CI: [ 0.104 ,  0.268 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-MC vs Gemini-2.0-Flash-MC
## ----------------------------------------
## Proportions:  0.573  vs  0.373
## Difference:  0.201
## Chi-squared:  23.4
## Degrees of freedom:  1
## P-value:  0.000001316
## 95% CI: [ 0.119 ,  0.282 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-MC vs Gemini-2.5-Pro-SC
## ----------------------------------------
## Proportions:  0.573  vs  0.467
## Difference:  0.106
## Chi-squared:  6.322
## Degrees of freedom:  1
## P-value:  0.01192
## 95% CI: [ 0.023 ,  0.189 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-MC vs Gemini-2.5-Pro-MC
## ----------------------------------------
## Proportions:  0.573  vs  0.456
## Difference:  0.117
## Chi-squared:  7.74
## Degrees of freedom:  1
## P-value:  0.005401
## 95% CI: [ 0.034 ,  0.2 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-MC vs ChatGPT-4o-SC
## ----------------------------------------
## Proportions:  0.573  vs  0.398
## Difference:  0.175
## Chi-squared:  17.732
## Degrees of freedom:  1
## P-value:  0.00002544
## 95% CI: [ 0.093 ,  0.257 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-MC vs ChatGPT-4o-MC
## ----------------------------------------
## Proportions:  0.573  vs  0.441
## Difference:  0.132
## Chi-squared:  9.936
## Degrees of freedom:  1
## P-value:  0.00162
## 95% CI: [ 0.049 ,  0.215 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-MC vs GPT-4.1-SC
## ----------------------------------------
## Proportions:  0.573  vs  0.441
## Difference:  0.132
## Chi-squared:  9.901
## Degrees of freedom:  1
## P-value:  0.001652
## 95% CI: [ 0.049 ,  0.214 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-MC vs GPT-4.1-MC
## ----------------------------------------
## Proportions:  0.573  vs  0.408
## Difference:  0.165
## Chi-squared:  15.692
## Degrees of freedom:  1
## P-value:  0.00007455
## 95% CI: [ 0.083 ,  0.247 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-MC vs GPT-4.1-GPT-Image-SC
## ----------------------------------------
## Proportions:  0.573  vs  0.386
## Difference:  0.187
## Chi-squared:  20.283
## Degrees of freedom:  1
## P-value:  0.000006678
## 95% CI: [ 0.105 ,  0.269 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-MC vs GPT-4.1-GPT-Image-MC
## ----------------------------------------
## Proportions:  0.573  vs  0.38
## Difference:  0.193
## Chi-squared:  21.679
## Degrees of freedom:  1
## P-value:  0.000003224
## 95% CI: [ 0.111 ,  0.275 ]
## Significant:  YES (p < 0.05)
##
##  Sonnet-4-SC vs Sonnet-4-MC
## ----------------------------------------
## Proportions:  0.428  vs  0.406
## Difference:  0.021
## Chi-squared:  0.203
## Degrees of freedom:  1
## P-value:  0.6521
## 95% CI: [ -0.061 ,  0.104 ]
## Significant:  NO
##
##  Sonnet-4-SC vs Gemini-2.0-Flash-SC
## ----------------------------------------
## Proportions:  0.428  vs  0.387
## Difference:  0.04
## Chi-squared:  0.851
## Degrees of freedom:  1
## P-value:  0.3562
## 95% CI: [ -0.042 ,  0.122 ]
## Significant:  NO
##
##  Sonnet-4-SC vs Gemini-2.0-Flash-MC
## ----------------------------------------
## Proportions:  0.428  vs  0.373
## Difference:  0.055
## Chi-squared:  1.668
## Degrees of freedom:  1
## P-value:  0.1965
## 95% CI: [ -0.027 ,  0.137 ]
## Significant:  NO
##
##  Sonnet-4-SC vs Gemini-2.5-Pro-SC
## ----------------------------------------
## Proportions:  0.428  vs  0.467
## Difference:  -0.04
## Chi-squared:  0.799
## Degrees of freedom:  1
## P-value:  0.3713
## 95% CI: [ -0.122 ,  0.043 ]
## Significant:  NO
##
##  Sonnet-4-SC vs Gemini-2.5-Pro-MC
## ----------------------------------------
## Proportions:  0.428  vs  0.456
## Difference:  -0.029
## Chi-squared:  0.39
## Degrees of freedom:  1
## P-value:  0.5321
## 95% CI: [ -0.111 ,  0.054 ]
## Significant:  NO
##
##  Sonnet-4-SC vs ChatGPT-4o-SC
## ----------------------------------------
## Proportions:  0.428  vs  0.398
## Difference:  0.03
## Chi-squared:  0.428
## Degrees of freedom:  1
## P-value:  0.5128
## 95% CI: [ -0.052 ,  0.112 ]
## Significant:  NO
##
##  Sonnet-4-SC vs ChatGPT-4o-MC
## ----------------------------------------
## Proportions:  0.428  vs  0.441
## Difference:  -0.014
## Chi-squared:  0.063
## Degrees of freedom:  1
## P-value:  0.8013
## 95% CI: [ -0.096 ,  0.069 ]
## Significant:  NO
##
##  Sonnet-4-SC vs GPT-4.1-SC
## ----------------------------------------
## Proportions:  0.428  vs  0.441
## Difference:  -0.014
## Chi-squared:  0.066
## Degrees of freedom:  1
## P-value:  0.7969
## 95% CI: [ -0.096 ,  0.069 ]
## Significant:  NO
##
##  Sonnet-4-SC vs GPT-4.1-MC
## ----------------------------------------
## Proportions:  0.428  vs  0.408
## Difference:  0.019
## Chi-squared:  0.161
## Degrees of freedom:  1
## P-value:  0.6882
## 95% CI: [ -0.063 ,  0.102 ]
## Significant:  NO
##
##  Sonnet-4-SC vs GPT-4.1-GPT-Image-SC
## ----------------------------------------
## Proportions:  0.428  vs  0.386
## Difference:  0.042
## Chi-squared:  0.906
## Degrees of freedom:  1
## P-value:  0.3411
## 95% CI: [ -0.04 ,  0.123 ]
## Significant:  NO
##
##  Sonnet-4-SC vs GPT-4.1-GPT-Image-MC
## ----------------------------------------
## Proportions:  0.428  vs  0.38
## Difference:  0.048
## Chi-squared:  1.225
## Degrees of freedom:  1
## P-value:  0.2683
## 95% CI: [ -0.034 ,  0.129 ]
## Significant:  NO
##
##  Sonnet-4-MC vs Gemini-2.0-Flash-SC
## ----------------------------------------
## Proportions:  0.406  vs  0.387
## Difference:  0.019
## Chi-squared:  0.151
## Degrees of freedom:  1
## P-value:  0.6975
## 95% CI: [ -0.063 ,  0.1 ]
## Significant:  NO
##
##  Sonnet-4-MC vs Gemini-2.0-Flash-MC
## ----------------------------------------
## Proportions:  0.406  vs  0.373
## Difference:  0.034
## Chi-squared:  0.575
## Degrees of freedom:  1
## P-value:  0.4484
## 95% CI: [ -0.048 ,  0.115 ]
## Significant:  NO
##
##  Sonnet-4-MC vs Gemini-2.5-Pro-SC
## ----------------------------------------
## Proportions:  0.406  vs  0.467
## Difference:  -0.061
## Chi-squared:  2.036
## Degrees of freedom:  1
## P-value:  0.1537
## 95% CI: [ -0.144 ,  0.021 ]
## Significant:  NO
##
##  Sonnet-4-MC vs Gemini-2.5-Pro-MC
## ----------------------------------------
## Proportions:  0.406  vs  0.456
## Difference:  -0.05
## Chi-squared:  1.34
## Degrees of freedom:  1
## P-value:  0.2469
## 95% CI: [ -0.133 ,  0.032 ]
## Significant:  NO
##
```

```
##  Sonnet-4-MC vs ChatGPT-4o-SC
## ----------------------------------------
## Proportions:  0.406  vs  0.398
## Difference:  0.008
## Chi-squared:  0.015
## Degrees of freedom:  1
## P-value:  0.9041
## 95% CI: [ -0.074 ,  0.09 ]
## Significant:  NO
##
##  Sonnet-4-MC vs ChatGPT-4o-MC
## ----------------------------------------
## Proportions:  0.406  vs  0.441
## Difference:  -0.035
## Chi-squared:  0.616
## Degrees of freedom:  1
## P-value:  0.4325
## 95% CI: [ -0.117 ,  0.047 ]
## Significant:  NO
##
##  Sonnet-4-MC vs GPT-4.1-SC
## ----------------------------------------
## Proportions:  0.406  vs  0.441
## Difference:  -0.035
## Chi-squared:  0.625
## Degrees of freedom:  1
## P-value:  0.4291
## 95% CI: [ -0.118 ,  0.047 ]
## Significant:  NO
##
##  Sonnet-4-MC vs GPT-4.1-MC
## ----------------------------------------
## Proportions:  0.406  vs  0.408
## Difference:  -0.002
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.083 ,  0.079 ]
## Significant:  NO
##
##  Sonnet-4-MC vs GPT-4.1-GPT-Image-SC
## ----------------------------------------
## Proportions:  0.406  vs  0.386
## Difference:  0.02
## Chi-squared:  0.175
## Degrees of freedom:  1
## P-value:  0.6758
## 95% CI: [ -0.062 ,  0.102 ]
## Significant:  NO
##
##  Sonnet-4-MC vs GPT-4.1-GPT-Image-MC
## ----------------------------------------
## Proportions:  0.406  vs  0.38
## Difference:  0.026
## Chi-squared:  0.329
## Degrees of freedom:  1
## P-value:  0.5665
## 95% CI: [ -0.055 ,  0.108 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-SC vs Gemini-2.0-Flash-MC
## ---------------------------------------
## Proportions:  0.387  vs  0.373
## Difference:  0.015
## Chi-squared:  0.082
## Degrees of freedom:  1
## P-value:  0.7751
## 95% CI: [ -0.066 ,  0.096 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-SC vs Gemini-2.5-Pro-SC
## ---------------------------------------
## Proportions:  0.387  vs  0.467
## Difference:  -0.08
## Chi-squared:  3.6
## Degrees of freedom:  1
## P-value:  0.05777
## 95% CI: [ -0.162 ,  0.002 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-SC vs Gemini-2.5-Pro-MC
## ---------------------------------------
## Proportions:  0.387  vs  0.456
## Difference:  -0.069
## Chi-squared:  2.653
## Degrees of freedom:  1
## P-value:  0.1033
## 95% CI: [ -0.151 ,  0.013 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-SC vs ChatGPT-4o-SC
## ---------------------------------------
## Proportions:  0.387  vs  0.398
## Difference:  -0.011
## Chi-squared:  0.034
## Degrees of freedom:  1
## P-value:  0.8534
## 95% CI: [ -0.092 ,  0.071 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-SC vs ChatGPT-4o-MC
## ---------------------------------------
## Proportions:  0.387  vs  0.441
## Difference:  -0.054
## Chi-squared:  1.579
## Degrees of freedom:  1
## P-value:  0.209
## 95% CI: [ -0.136 ,  0.028 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-SC vs GPT-4.1-SC
## ----------------------------------------
## Proportions:  0.387  vs  0.441
## Difference:  -0.054
## Chi-squared:  1.593
## Degrees of freedom:  1
## P-value:  0.2069
## 95% CI: [ -0.136 ,  0.028 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-SC vs GPT-4.1-MC
## ----------------------------------------
## Proportions:  0.387  vs  0.408
## Difference:  -0.021
## Chi-squared:  0.192
## Degrees of freedom:  1
## P-value:  0.6612
## 95% CI: [ -0.102 ,  0.061 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-SC vs GPT-4.1-GPT-Image-SC
## ----------------------------------------
## Proportions:  0.387  vs  0.386
## Difference:  0.001
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.078 ,  0.08 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-SC vs GPT-4.1-GPT-Image-MC
## ----------------------------------------
## Proportions:  0.387  vs  0.38
## Difference:  0.007
## Chi-squared:  0.01
## Degrees of freedom:  1
## P-value:  0.9198
## 95% CI: [ -0.074 ,  0.088 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-MC vs Gemini-2.5-Pro-SC
## ----------------------------------------
## Proportions:  0.373  vs  0.467
## Difference:  -0.095
## Chi-squared:  5.132
## Degrees of freedom:  1
## P-value:  0.02349
## 95% CI: [ -0.177 ,  -0.013 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-MC vs Gemini-2.5-Pro-MC
## ----------------------------------------
## Proportions:  0.373  vs  0.456
## Difference:  -0.084
## Chi-squared:  3.989
## Degrees of freedom:  1
## P-value:  0.0458
## 95% CI: [ -0.166 ,  -0.002 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-MC vs ChatGPT-4o-SC
## ----------------------------------------
## Proportions:  0.373  vs  0.398
## Difference:  -0.025
## Chi-squared:  0.307
## Degrees of freedom:  1
## P-value:  0.5794
## 95% CI: [ -0.107 ,  0.056 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-MC vs ChatGPT-4o-MC
## ----------------------------------------
## Proportions:  0.373  vs  0.441
## Difference:  -0.069
## Chi-squared:  2.641
## Degrees of freedom:  1
## P-value:  0.1041
## 95% CI: [ -0.15 ,  0.013 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-MC vs GPT-4.1-SC
## ----------------------------------------
## Proportions:  0.373  vs  0.441
## Difference:  -0.069
## Chi-squared:  2.66
## Degrees of freedom:  1
## P-value:  0.1029
## 95% CI: [ -0.151 ,  0.013 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-MC vs GPT-4.1-MC
## ----------------------------------------
## Proportions:  0.373  vs  0.408
## Difference:  -0.036
## Chi-squared:  0.652
## Degrees of freedom:  1
## P-value:  0.4193
## 95% CI: [ -0.117 ,  0.046 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-MC vs GPT-4.1-GPT-Image-SC
## ----------------------------------------
## Proportions:  0.373  vs  0.386
## Difference:  -0.013
## Chi-squared:  0.066
## Degrees of freedom:  1
## P-value:  0.7978
## 95% CI: [ -0.094 ,  0.067 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-MC vs GPT-4.1-GPT-Image-MC
## ----------------------------------------
## Proportions:  0.373  vs  0.38
## Difference:  -0.007
## Chi-squared:  0.01
## Degrees of freedom:  1
## P-value:  0.9197
## 95% CI: [ -0.088 ,  0.074 ]
## Significant:  NO
##
##  Gemini-2.5-Pro-SC vs Gemini-2.5-Pro-MC
## ----------------------------------------
## Proportions:  0.467  vs  0.456
## Difference:  0.011
## Chi-squared:  0.035
## Degrees of freedom:  1
## P-value:  0.8514
## 95% CI: [ -0.072 ,  0.094 ]
## Significant:  NO
##
##  Gemini-2.5-Pro-SC vs ChatGPT-4o-SC
## ----------------------------------------
## Proportions:  0.467  vs  0.398
## Difference:  0.069
## Chi-squared:  2.657
## Degrees of freedom:  1
## P-value:  0.1031
## 95% CI: [ -0.013 ,  0.152 ]
## Significant:  NO
##
##  Gemini-2.5-Pro-SC vs ChatGPT-4o-MC
## ----------------------------------------
## Proportions:  0.467  vs  0.441
## Difference:  0.026
## Chi-squared:  0.314
## Degrees of freedom:  1
## P-value:  0.5753
## 95% CI: [ -0.057 ,  0.109 ]
## Significant:  NO
##
##  Gemini-2.5-Pro-SC vs GPT-4.1-SC
## ----------------------------------------
## Proportions:  0.467  vs  0.441
## Difference:  0.026
## Chi-squared:  0.307
## Degrees of freedom:  1
## P-value:  0.5792
## 95% CI: [ -0.057 ,  0.109 ]
## Significant:  NO
##
```

```
##  Gemini-2.5-Pro-SC vs GPT-4.1-MC
## ----------------------------------------
## Proportions:  0.467  vs  0.408
## Difference:  0.059
## Chi-squared:  1.897
## Degrees of freedom:  1
## P-value:  0.1684
## 95% CI: [ -0.023 ,  0.142 ]
## Significant:  NO
##
##  Gemini-2.5-Pro-SC vs GPT-4.1-GPT-Image-SC
## ----------------------------------------
## Proportions:  0.467  vs  0.386
## Difference:  0.081
## Chi-squared:  3.712
## Degrees of freedom:  1
## P-value:  0.05401
## 95% CI: [ -0.001 ,  0.163 ]
## Significant:  NO
##
##  Gemini-2.5-Pro-SC vs GPT-4.1-GPT-Image-MC
## ----------------------------------------
## Proportions:  0.467  vs  0.38
## Difference:  0.087
## Chi-squared:  4.332
## Degrees of freedom:  1
## P-value:  0.0374
## 95% CI: [ 0.005 ,  0.169 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5-Pro-MC vs ChatGPT-4o-SC
## ----------------------------------------
## Proportions:  0.456  vs  0.398
## Difference:  0.058
## Chi-squared:  1.853
## Degrees of freedom:  1
## P-value:  0.1735
## 95% CI: [ -0.024 ,  0.141 ]
## Significant:  NO
##
##  Gemini-2.5-Pro-MC vs ChatGPT-4o-MC
## ----------------------------------------
## Proportions:  0.456  vs  0.441
## Difference:  0.015
## Chi-squared:  0.085
## Degrees of freedom:  1
## P-value:  0.7711
## 95% CI: [ -0.068 ,  0.098 ]
## Significant:  NO
##
##  Gemini-2.5-Pro-MC vs GPT-4.1-SC
## ----------------------------------------
## Proportions:  0.456  vs  0.441
## Difference:  0.015
## Chi-squared:  0.081
## Degrees of freedom:  1
## P-value:  0.7755
## 95% CI: [ -0.068 ,  0.098 ]
## Significant:  NO
##
##  Gemini-2.5-Pro-MC vs GPT-4.1-MC
## ----------------------------------------
## Proportions:  0.456  vs  0.408
## Difference:  0.048
## Chi-squared:  1.228
## Degrees of freedom:  1
## P-value:  0.2677
## 95% CI: [ -0.034 ,  0.131 ]
## Significant:  NO
##
##  Gemini-2.5-Pro-MC vs GPT-4.1-GPT-Image-SC
## ----------------------------------------
## Proportions:  0.456  vs  0.386
## Difference:  0.07
## Chi-squared:  2.75
## Degrees of freedom:  1
## P-value:  0.09728
## 95% CI: [ -0.012 ,  0.152 ]
## Significant:  NO
##
##  Gemini-2.5-Pro-MC vs GPT-4.1-GPT-Image-MC
## ----------------------------------------
## Proportions:  0.456  vs  0.38
## Difference:  0.076
## Chi-squared:  3.287
## Degrees of freedom:  1
## P-value:  0.06985
## 95% CI: [ -0.006 ,  0.158 ]
## Significant:  NO
##
##  ChatGPT-4o-SC vs ChatGPT-4o-MC
## ----------------------------------------
## Proportions:  0.398  vs  0.441
## Difference:  -0.043
## Chi-squared:  0.977
## Degrees of freedom:  1
## P-value:  0.3229
## 95% CI: [ -0.125 ,  0.039 ]
## Significant:  NO
##
##  ChatGPT-4o-SC vs GPT-4.1-SC
## ----------------------------------------
## Proportions:  0.398  vs  0.441
## Difference:  -0.043
## Chi-squared:  0.988
## Degrees of freedom:  1
## P-value:  0.3201
## 95% CI: [ -0.126 ,  0.039 ]
## Significant:  NO
##
##  ChatGPT-4o-SC vs GPT-4.1-MC
## ----------------------------------------
## Proportions:  0.398  vs  0.408
## Difference:  -0.01
## Chi-squared:  0.029
## Degrees of freedom:  1
## P-value:  0.8649
## 95% CI: [ -0.092 ,  0.072 ]
## Significant:  NO
##
##  ChatGPT-4o-SC vs GPT-4.1-GPT-Image-SC
## ----------------------------------------
## Proportions:  0.398  vs  0.386
## Difference:  0.012
## Chi-squared:  0.046
## Degrees of freedom:  1
## P-value:  0.8304
## 95% CI: [ -0.07 ,  0.093 ]
## Significant:  NO
##
##  ChatGPT-4o-SC vs GPT-4.1-GPT-Image-MC
## ----------------------------------------
## Proportions:  0.398  vs  0.38
## Difference:  0.018
## Chi-squared:  0.136
## Degrees of freedom:  1
## P-value:  0.7119
## 95% CI: [ -0.063 ,  0.099 ]
## Significant:  NO
##
##  ChatGPT-4o-MC vs GPT-4.1-SC
## ----------------------------------------
## Proportions:  0.441  vs  0.441
## Difference:  0
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.08 ,  0.079 ]
## Significant:  NO
##
##  ChatGPT-4o-MC vs GPT-4.1-MC
## ----------------------------------------
## Proportions:  0.441  vs  0.408
## Difference:  0.033
## Chi-squared:  0.541
## Degrees of freedom:  1
## P-value:  0.4621
## 95% CI: [ -0.049 ,  0.115 ]
## Significant:  NO
##
##  ChatGPT-4o-MC vs GPT-4.1-GPT-Image-SC
## ----------------------------------------
## Proportions:  0.441  vs  0.386
## Difference:  0.055
## Chi-squared:  1.653
## Degrees of freedom:  1
## P-value:  0.1985
## 95% CI: [ -0.027 ,  0.137 ]
## Significant:  NO
##
##  ChatGPT-4o-MC vs GPT-4.1-GPT-Image-MC
## ----------------------------------------
## Proportions:  0.441  vs  0.38
## Difference:  0.061
## Chi-squared:  2.075
## Degrees of freedom:  1
## P-value:  0.1497
## 95% CI: [ -0.021 ,  0.143 ]
## Significant:  NO
##
##  GPT-4.1-SC vs GPT-4.1-MC
## ----------------------------------------
## Proportions:  0.441  vs  0.408
## Difference:  0.033
## Chi-squared:  0.549
## Degrees of freedom:  1
## P-value:  0.4586
## 95% CI: [ -0.049 ,  0.116 ]
## Significant:  NO
##
##  GPT-4.1-SC vs GPT-4.1-GPT-Image-SC
## ----------------------------------------
## Proportions:  0.441  vs  0.386
## Difference:  0.055
## Chi-squared:  1.668
## Degrees of freedom:  1
## P-value:  0.1966
## 95% CI: [ -0.027 ,  0.137 ]
## Significant:  NO
##
##  GPT-4.1-SC vs GPT-4.1-GPT-Image-MC
## ----------------------------------------
## Proportions:  0.441  vs  0.38
## Difference:  0.061
## Chi-squared:  2.092
## Degrees of freedom:  1
## P-value:  0.1481
## 95% CI: [ -0.02 ,  0.143 ]
## Significant:  NO
##
##  GPT-4.1-MC vs GPT-4.1-GPT-Image-SC
## ----------------------------------------
## Proportions:  0.408  vs  0.386
## Difference:  0.022
## Chi-squared:  0.219
## Degrees of freedom:  1
## P-value:  0.64
## 95% CI: [ -0.06 ,  0.104 ]
## Significant:  NO
##
##  GPT-4.1-MC vs GPT-4.1-GPT-Image-MC
## ----------------------------------------
## Proportions:  0.408  vs  0.38
## Difference:  0.028
## Chi-squared:  0.388
## Degrees of freedom:  1
## P-value:  0.5334
## 95% CI: [ -0.053 ,  0.11 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image-SC vs GPT-4.1-GPT-Image-MC
## ----------------------------------------
## Proportions:  0.386  vs  0.38
## Difference:  0.006
## Chi-squared:  0.005
## Degrees of freedom:  1
## P-value:  0.9432
## 95% CI: [ -0.075 ,  0.087 ]
## Significant:  NO
```

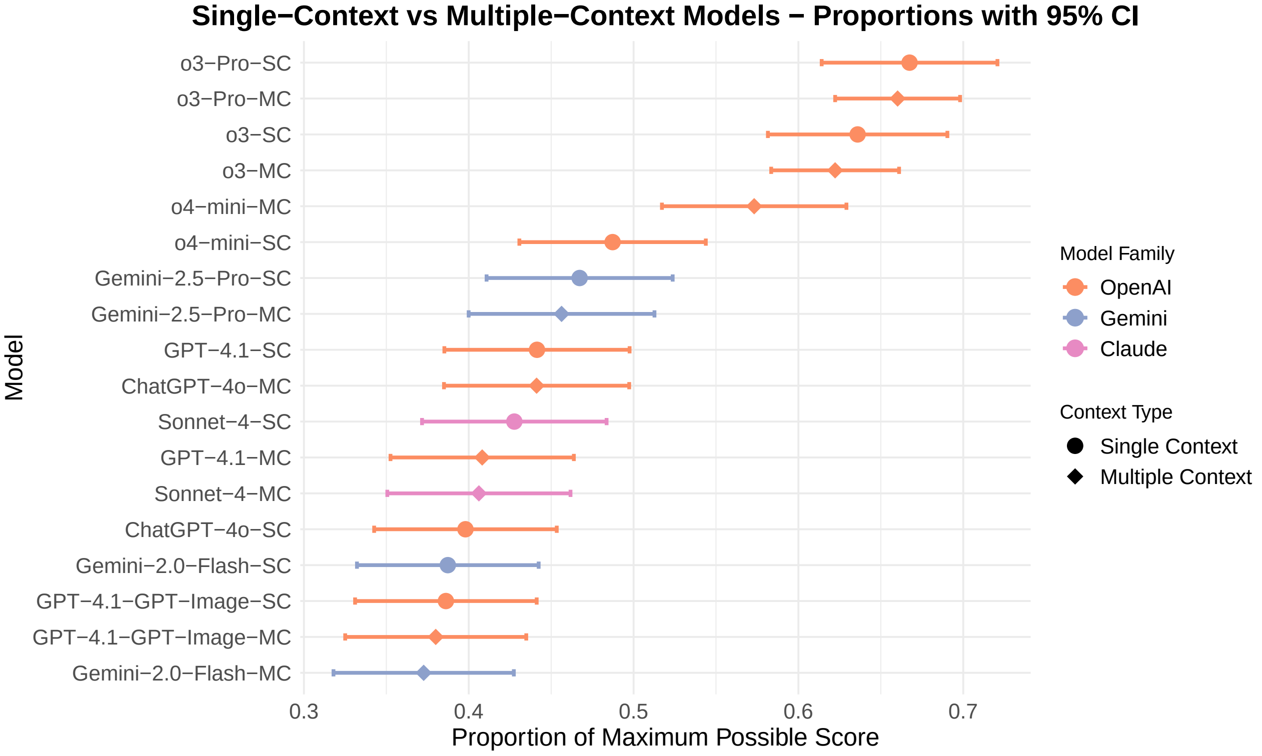
```
##
##
## Summary Table - Single-Context vs Multiple-Context:
```

```
##
##
##                comparison                                        diff   chi_squared   p_value  signi
## -------------  ---------------------------------------------  -------  ------------  --------  -----
## X-squared      o3-SC vs o3-MC                                   0.014     0.1064260    0.7443  FALSE
## X-squared1     o3-SC vs o3-Pro-SC                              -0.032     0.5244141    0.4690  FALSE
## X-squared2     o3-SC vs o3-Pro-MC                              -0.024     0.4171953    0.5183  FALSE
## X-squared3     o3-SC vs o4-mini-SC                              0.149    12.8683810    0.0003  TRUE
## X-squared4     o3-SC vs o4-mini-MC                              0.063     2.2159433    0.1366  FALSE
## X-squared5     o3-SC vs Sonnet-4-SC                             0.208    25.3065478    0.0000  TRUE
## X-squared6     o3-SC vs Sonnet-4-MC                             0.230    30.8197762    0.0000  TRUE
## X-squared7     o3-SC vs Gemini-2.0-Flash-SC                     0.249    36.1232148    0.0000  TRUE
## X-squared8     o3-SC vs Gemini-2.0-Flash-MC                     0.263    40.5493640    0.0000  TRUE
## X-squared9     o3-SC vs Gemini-2.5-Pro-SC                       0.169    16.5768348    0.0000  TRUE
## X-squared10    o3-SC vs Gemini-2.5-Pro-MC                       0.180    18.8068512    0.0000  TRUE
## X-squared11    o3-SC vs ChatGPT-4o-SC                           0.238    33.0601386    0.0000  TRUE
## X-squared12    o3-SC vs ChatGPT-4o-MC                           0.195    22.1202375    0.0000  TRUE
## X-squared13    o3-SC vs GPT-4.1-SC                              0.195    22.0678694    0.0000  TRUE
## X-squared14    o3-SC vs GPT-4.1-MC                              0.228    30.2861685    0.0000  TRUE
## X-squared15    o3-SC vs GPT-4.1-GPT-Image-SC                    0.250    36.4668794    0.0000  TRUE
## X-squared16    o3-SC vs GPT-4.1-GPT-Image-MC                    0.256    38.3043217    0.0000  TRUE
## X-squared17    o3-MC vs o3-Pro-SC                              -0.045     1.5747438    0.2095  FALSE
## X-squared18    o3-MC vs o3-Pro-MC                              -0.038     1.7130313    0.1906  FALSE
## X-squared19    o3-MC vs o4-mini-SC                              0.135    14.3908156    0.0001  TRUE
## X-squared20    o3-MC vs o4-mini-MC                              0.049     1.8192544    0.1774  FALSE
## X-squared21    o3-MC vs Sonnet-4-SC                             0.195    29.9264284    0.0000  TRUE
## X-squared22    o3-MC vs Sonnet-4-MC                             0.216    36.8780816    0.0000  TRUE
## X-squared23    o3-MC vs Gemini-2.0-Flash-SC                     0.235    43.5735443    0.0000  TRUE
## X-squared24    o3-MC vs Gemini-2.0-Flash-MC                     0.250    49.1598340    0.0000  TRUE
## X-squared25    o3-MC vs Gemini-2.5-Pro-SC                       0.155    18.9850421    0.0000  TRUE
## X-squared26    o3-MC vs Gemini-2.5-Pro-MC                       0.166    21.7672906    0.0000  TRUE
## X-squared27    o3-MC vs ChatGPT-4o-SC                           0.224    39.7062102    0.0000  TRUE
## X-squared28    o3-MC vs ChatGPT-4o-MC                           0.181    25.9194238    0.0000  TRUE
## X-squared29    o3-MC vs GPT-4.1-SC                              0.181    25.8536662    0.0000  TRUE
## X-squared30    o3-MC vs GPT-4.1-MC                              0.214    36.2046458    0.0000  TRUE
## X-squared31    o3-MC vs GPT-4.1-GPT-Image-SC                    0.236    44.0074173    0.0000  TRUE
## X-squared32    o3-MC vs GPT-4.1-GPT-Image-MC                    0.242    46.3268747    0.0000  TRUE
## X-squared33    o3-Pro-SC vs o3-Pro-MC                           0.007     0.0201847    0.8870  FALSE
## X-squared34    o3-Pro-SC vs o4-mini-SC                          0.180    19.2229175    0.0000  TRUE
## X-squared35    o3-Pro-SC vs o4-mini-MC                          0.094     5.2661001    0.0217  TRUE
## X-squared36    o3-Pro-SC vs Sonnet-4-SC                         0.240    33.8541333    0.0000  TRUE
## X-squared37    o3-Pro-SC vs Sonnet-4-MC                         0.261    40.1386029    0.0000  TRUE
## X-squared38    o3-Pro-SC vs Gemini-2.0-Flash-SC                 0.280    46.1109296    0.0000  TRUE
## X-squared39    o3-Pro-SC vs Gemini-2.0-Flash-MC                 0.295    51.0514768    0.0000  TRUE
## X-squared40    o3-Pro-SC vs Gemini-2.5-Pro-SC                   0.200    23.6756299    0.0000  TRUE
## X-squared41    o3-Pro-SC vs Gemini-2.5-Pro-MC                   0.211    26.3098049    0.0000  TRUE
## X-squared42    o3-Pro-SC vs ChatGPT-4o-SC                       0.269    42.6692467    0.0000  TRUE
## X-squared43    o3-Pro-SC vs ChatGPT-4o-MC                       0.226    30.1773792    0.0000  TRUE
## X-squared44    o3-Pro-SC vs GPT-4.1-SC                          0.226    30.1166275    0.0000  TRUE
## X-squared45    o3-Pro-SC vs GPT-4.1-MC                          0.259    39.5340433    0.0000  TRUE
## X-squared46    o3-Pro-SC vs GPT-4.1-GPT-Image-SC                0.281    46.4958563    0.0000  TRUE
## X-squared47    o3-Pro-SC vs GPT-4.1-GPT-Image-MC                0.287    48.5500288    0.0000  TRUE
## X-squared48    o3-Pro-MC vs o4-mini-SC                          0.173    24.2547825    0.0000  TRUE
## X-squared49    o3-Pro-MC vs o4-mini-MC                          0.087     6.1367646    0.0132  TRUE
## X-squared50    o3-Pro-MC vs Sonnet-4-SC                         0.233    43.5260055    0.0000  TRUE
## X-squared51    o3-Pro-MC vs Sonnet-4-MC                         0.254    51.7942706    0.0000  TRUE
## X-squared52    o3-Pro-MC vs Gemini-2.0-Flash-SC                 0.273    59.6311060    0.0000  TRUE
## X-squared53    o3-Pro-MC vs Gemini-2.0-Flash-MC                 0.288    66.0942845    0.0000  TRUE
## X-squared54    o3-Pro-MC vs Gemini-2.5-Pro-SC                   0.193    30.1141832    0.0000  TRUE
## X-squared55    o3-Pro-MC vs Gemini-2.5-Pro-MC                   0.204    33.5848305    0.0000  TRUE
## X-squared56    o3-Pro-MC vs ChatGPT-4o-SC                       0.262    55.1178401    0.0000  TRUE
## X-squared57    o3-Pro-MC vs ChatGPT-4o-MC                       0.219    38.6819968    0.0000  TRUE
## X-squared58    o3-Pro-MC vs GPT-4.1-SC                          0.219    38.6019359    0.0000  TRUE
## X-squared59    o3-Pro-MC vs GPT-4.1-MC                          0.252    50.9997192    0.0000  TRUE
## X-squared60    o3-Pro-MC vs GPT-4.1-GPT-Image-SC                0.274    60.1353518    0.0000  TRUE
## X-squared61    o3-Pro-MC vs GPT-4.1-GPT-Image-MC                0.280    62.8243540    0.0000  TRUE
## X-squared62    o4-mini-SC vs o4-mini-MC                        -0.086     4.1058213    0.0427  TRUE
## X-squared63    o4-mini-SC vs Sonnet-4-SC                        0.060     1.9149825    0.1664  FALSE
## X-squared64    o4-mini-SC vs Sonnet-4-MC                        0.081     3.6709551    0.0554  FALSE
## X-squared65    o4-mini-SC vs Gemini-2.0-Flash-SC                0.100     5.6925532    0.0170  TRUE
## X-squared66    o4-mini-SC vs Gemini-2.0-Flash-MC                0.115     7.5799978    0.0059  TRUE
## X-squared67    o4-mini-SC vs Gemini-2.5-Pro-SC                  0.020     0.1669324    0.6829  FALSE
## X-squared68    o4-mini-SC vs Gemini-2.5-Pro-MC                  0.031     0.4592808    0.4980  FALSE
## X-squared69    o4-mini-SC vs ChatGPT-4o-SC                      0.089     4.4897434    0.0341  TRUE
## X-squared70    o4-mini-SC vs ChatGPT-4o-MC                      0.046     1.1032067    0.2936  FALSE
## X-squared71    o4-mini-SC vs GPT-4.1-SC                         0.046     1.0913292    0.2962  FALSE
## X-squared72    o4-mini-SC vs GPT-4.1-MC                         0.079     3.4841966    0.0620  FALSE
## X-squared73    o4-mini-SC vs GPT-4.1-GPT-Image-SC               0.101     5.8330562    0.0157  TRUE
## X-squared74    o4-mini-SC vs GPT-4.1-GPT-Image-MC               0.107     6.6020180    0.0102  TRUE
## X-squared75    o4-mini-MC vs Sonnet-4-SC                        0.146    12.1315640    0.0005  TRUE
## X-squared76    o4-mini-MC vs Sonnet-4-MC                        0.167    16.0812760    0.0001  TRUE
## X-squared77    o4-mini-MC vs Gemini-2.0-Flash-SC               0.186    20.0237337    0.0000  TRUE
## X-squared78    o4-mini-MC vs Gemini-2.0-Flash-MC               0.201    23.4000962    0.0000  TRUE
## X-squared79    o4-mini-MC vs Gemini-2.5-Pro-SC                 0.106     6.3223902    0.0119  TRUE
## X-squared80    o4-mini-MC vs Gemini-2.5-Pro-MC                 0.117     7.7398604    0.0054  TRUE
## X-squared81    o4-mini-MC vs ChatGPT-4o-SC                     0.175    17.7315722    0.0000  TRUE
## X-squared82    o4-mini-MC vs ChatGPT-4o-MC                     0.132     9.9363335    0.0016  TRUE
## X-squared83    o4-mini-MC vs GPT-4.1-SC                        0.132     9.9008851    0.0017  TRUE
## X-squared84    o4-mini-MC vs GPT-4.1-MC                        0.165    15.6917631    0.0001  TRUE
## X-squared85    o4-mini-MC vs GPT-4.1-GPT-Image-SC              0.187    20.2832914    0.0000  TRUE
## X-squared86    o4-mini-MC vs GPT-4.1-GPT-Image-MC              0.193    21.6786709    0.0000  TRUE
## X-squared87    Sonnet-4-SC vs Sonnet-4-MC                      0.021     0.2032859    0.6521  FALSE
## X-squared88    Sonnet-4-SC vs Gemini-2.0-Flash-SC              0.040     0.8510755    0.3562  FALSE
## X-squared89    Sonnet-4-SC vs Gemini-2.0-Flash-MC              0.055     1.6682043    0.1965  FALSE
## X-squared90    Sonnet-4-SC vs Gemini-2.5-Pro-SC               -0.040     0.7991629    0.3713  FALSE
## X-squared91    Sonnet-4-SC vs Gemini-2.5-Pro-MC               -0.029     0.3903232    0.5321  FALSE
## X-squared92    Sonnet-4-SC vs ChatGPT-4o-SC                    0.030     0.4283362    0.5128  FALSE
## X-squared93    Sonnet-4-SC vs ChatGPT-4o-MC                   -0.014     0.0633493    0.8013  FALSE
## X-squared94    Sonnet-4-SC vs GPT-4.1-SC                      -0.014     0.0662403    0.7969  FALSE
## X-squared95    Sonnet-4-SC vs GPT-4.1-MC                       0.019     0.1610833    0.6882  FALSE
## X-squared96    Sonnet-4-SC vs GPT-4.1-GPT-Image-SC             0.042     0.9061713    0.3411  FALSE
## X-squared97    Sonnet-4-SC vs GPT-4.1-GPT-Image-MC             0.048     1.2252528    0.2683  FALSE
## X-squared98    Sonnet-4-MC vs Gemini-2.0-Flash-SC              0.019     0.1511137    0.6975  FALSE
## X-squared99    Sonnet-4-MC vs Gemini-2.0-Flash-MC              0.034     0.5747113    0.4484  FALSE
## X-squared100   Sonnet-4-MC vs Gemini-2.5-Pro-SC               -0.061     2.0355735    0.1537  FALSE
## X-squared101   Sonnet-4-MC vs Gemini-2.5-Pro-MC               -0.050     1.3404799    0.2469  FALSE
## X-squared102   Sonnet-4-MC vs ChatGPT-4o-SC                    0.008     0.0145279    0.9041  FALSE
## X-squared103   Sonnet-4-MC vs ChatGPT-4o-MC                   -0.035     0.6161811    0.4325  FALSE
## X-squared104   Sonnet-4-MC vs GPT-4.1-SC                      -0.035     0.6251246    0.4291  FALSE
## X-squared105   Sonnet-4-MC vs GPT-4.1-MC                      -0.002     0.0000000    1.0000  FALSE
## X-squared106   Sonnet-4-MC vs GPT-4.1-GPT-Image-SC             0.020     0.1748460    0.6758  FALSE
## X-squared107   Sonnet-4-MC vs GPT-4.1-GPT-Image-MC             0.026     0.3286100    0.5665  FALSE
## X-squared108   Gemini-2.0-Flash-SC vs Gemini-2.0-Flash-MC      0.015     0.0816078    0.7751  FALSE
## X-squared109   Gemini-2.0-Flash-SC vs Gemini-2.5-Pro-SC       -0.080     3.6002593    0.0578  FALSE
## X-squared110   Gemini-2.0-Flash-SC vs Gemini-2.5-Pro-MC       -0.069     2.6530962    0.1033  FALSE
## X-squared111   Gemini-2.0-Flash-SC vs ChatGPT-4o-SC           -0.011     0.0341562    0.8534  FALSE
## X-squared112   Gemini-2.0-Flash-SC vs ChatGPT-4o-MC           -0.054     1.5785238    0.2090  FALSE
## X-squared113   Gemini-2.0-Flash-SC vs GPT-4.1-SC              -0.054     1.5928066    0.2069  FALSE
## X-squared114   Gemini-2.0-Flash-SC vs GPT-4.1-MC              -0.021     0.1921225    0.6612  FALSE
## X-squared115   Gemini-2.0-Flash-SC vs GPT-4.1-GPT-Image-SC     0.001     0.0000000    1.0000  FALSE
## X-squared116   Gemini-2.0-Flash-SC vs GPT-4.1-GPT-Image-MC     0.007     0.0101498    0.9198  FALSE
## X-squared117   Gemini-2.0-Flash-MC vs Gemini-2.5-Pro-SC       -0.095     5.1322474    0.0235  TRUE
## X-squared118   Gemini-2.0-Flash-MC vs Gemini-2.5-Pro-MC       -0.084     3.9887672    0.0458  TRUE
## X-squared119   Gemini-2.0-Flash-MC vs ChatGPT-4o-SC           -0.025     0.3071885    0.5794  FALSE
## X-squared120   Gemini-2.0-Flash-MC vs ChatGPT-4o-MC           -0.069     2.6411322    0.1041  FALSE
## X-squared121   Gemini-2.0-Flash-MC vs GPT-4.1-SC              -0.069     2.6595797    0.1029  FALSE
## X-squared122   Gemini-2.0-Flash-MC vs GPT-4.1-MC              -0.036     0.6523157    0.4193  FALSE
## X-squared123   Gemini-2.0-Flash-MC vs GPT-4.1-GPT-Image-SC    -0.013     0.0656431    0.7978  FALSE
## X-squared124   Gemini-2.0-Flash-MC vs GPT-4.1-GPT-Image-MC    -0.007     0.0101651    0.9197  FALSE
## X-squared125   Gemini-2.5-Pro-SC vs Gemini-2.5-Pro-MC          0.011     0.0351076    0.8514  FALSE
## X-squared126   Gemini-2.5-Pro-SC vs ChatGPT-4o-SC              0.069     2.6568007    0.1031  FALSE
## X-squared127   Gemini-2.5-Pro-SC vs ChatGPT-4o-MC              0.026     0.3138191    0.5753  FALSE
## X-squared128   Gemini-2.5-Pro-SC vs GPT-4.1-SC                 0.026     0.3074950    0.5792  FALSE
## X-squared129   Gemini-2.5-Pro-SC vs GPT-4.1-MC                 0.059     1.8969356    0.1684  FALSE
## X-squared130   Gemini-2.5-Pro-SC vs GPT-4.1-GPT-Image-SC       0.081     3.7123815    0.0540  FALSE
## X-squared131   Gemini-2.5-Pro-SC vs GPT-4.1-GPT-Image-MC        0.087     4.3318957    0.0374  TRUE
## X-squared132   Gemini-2.5-Pro-MC vs ChatGPT-4o-SC               0.058     1.8527550    0.1735  FALSE
## X-squared133   Gemini-2.5-Pro-MC vs ChatGPT-4o-MC               0.015     0.0846247    0.7711  FALSE
## X-squared134   Gemini-2.5-Pro-MC vs GPT-4.1-SC                  0.015     0.0813555    0.7755  FALSE
## X-squared135   Gemini-2.5-Pro-MC vs GPT-4.1-MC                  0.048     1.2283709    0.2677  FALSE
## X-squared136   Gemini-2.5-Pro-MC vs GPT-4.1-GPT-Image-SC        0.070     2.7495509    0.0973  FALSE
## X-squared137   Gemini-2.5-Pro-MC vs GPT-4.1-GPT-Image-MC        0.076     3.2865429    0.0698  FALSE
## X-squared138   ChatGPT-4o-SC vs ChatGPT-4o-MC                  -0.043     0.9771033    0.3229  FALSE
## X-squared139   ChatGPT-4o-SC vs GPT-4.1-SC                     -0.043     0.9883535    0.3201  FALSE
## X-squared140   ChatGPT-4o-SC vs GPT-4.1-MC                     -0.010     0.0289418    0.8649  FALSE
## X-squared141   ChatGPT-4o-SC vs GPT-4.1-GPT-Image-SC            0.012     0.0458944    0.8304  FALSE
## X-squared142   ChatGPT-4o-SC vs GPT-4.1-GPT-Image-MC            0.018     0.1364216    0.7119  FALSE
## X-squared143   ChatGPT-4o-MC vs GPT-4.1-SC                      0.000     0.0000000    1.0000  FALSE
## X-squared144   ChatGPT-4o-MC vs GPT-4.1-MC                      0.033     0.5409148    0.4621  FALSE
## X-squared145   ChatGPT-4o-MC vs GPT-4.1-GPT-Image-SC            0.055     1.6531946    0.1985  FALSE
## X-squared146   ChatGPT-4o-MC vs GPT-4.1-GPT-Image-MC            0.061     2.0754248    0.1497  FALSE
## X-squared147   GPT-4.1-SC vs GPT-4.1-MC                         0.033     0.5492969    0.4586  FALSE
## X-squared148   GPT-4.1-SC vs GPT-4.1-GPT-Image-SC               0.055     1.6678095    0.1966  FALSE
## X-squared149   GPT-4.1-SC vs GPT-4.1-GPT-Image-MC               0.061     2.0917899    0.1481  FALSE
## X-squared150   GPT-4.1-MC vs GPT-4.1-GPT-Image-SC               0.022     0.2187704    0.6400  FALSE
## X-squared151   GPT-4.1-MC vs GPT-4.1-GPT-Image-MC               0.028     0.3879056    0.5334  FALSE
## X-squared152   GPT-4.1-GPT-Image-SC vs GPT-4.1-GPT-Image-MC     0.006     0.0050812    0.9432  FALSE
```

Visualization of Single vs. Multiple Context

```r
# Plot for Single-Context vs Multiple-Context
sc_mc_plot <- ggplot(sc_mc_data, aes(x = reorder(model, proportion), y = proportion, color = model)) +
  geom_point(size = 4, aes(color = as.factor(color), shape = as.factor(shape))) +
  geom_errorbar(aes(ymin = proportion - 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    ymax = proportion + 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    color = color),
                width = 0.2, size = 1) +
  coord_flip() +
  theme_minimal() +
  labs(title = "Single-Context vs Multiple-Context Models - Proportions with 95% CI",
       x = "Model",
       y = "Proportion of Maximum Possible Score") +
  theme(plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
        axis.text = element_text(size = 12),
        axis.title = element_text(size = 14),
        legend.text = element_text(size = 12)) +
  scale_color_manual(
    values = c("#fc8d62", "#8da0cb", "#e78ac3"),
    name = "Model Family",
    breaks = c("#fc8d62", "#8da0cb", "#e78ac3"),
    labels = c("OpenAI", "Gemini", "Claude")
  ) +
  scale_shape_manual(
    values = c(16, 18),
    name = "Context Type",
    breaks = c(16, 18),
    labels = c("Single Context", "Multiple Context")
  )
```

```r
print(sc_mc_plot)
```



Finke et al. Tasks - All Pairwise Comparisons

```r
# 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:
```

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

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

```r
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:  0.63  vs  0.611
## Difference:  0.02
## Chi-squared:  0.189
## Degrees of freedom:  1
## P-value:  0.6636
## 95% CI: [ -0.059 ,  0.098 ]
## Significant:  NO
##
##  Humans vs o3-GPT-Image
## ----------------------------------------
## Proportions:  0.63  vs  0.56
## Difference:  0.07
## Chi-squared:  4.009
## Degrees of freedom:  1
## P-value:  0.04525
## 95% CI: [ 0 ,  0.14 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o3-Pro
## ----------------------------------------
## Proportions:  0.63  vs  0.772
## Difference:  -0.141
## Chi-squared:  13.467
## Degrees of freedom:  1
## P-value:  0.0002428
## 95% CI: [ -0.211 ,  -0.072 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-4.1
## ----------------------------------------
## Proportions:  0.63  vs  0.47
## Difference:  0.16
## Chi-squared:  11.426
## Degrees of freedom:  1
## P-value:  0.0007242
## 95% CI: [ 0.063 ,  0.257 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.63  vs  0.342
## Difference:  0.289
## Chi-squared:  37.705
## Degrees of freedom:  1
## P-value:  0.0000000008229
## 95% CI: [ 0.196 ,  0.381 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.63  vs  0.408
## Difference:  0.222
## Chi-squared:  22.217
## Degrees of freedom:  1
## P-value:  0.000002435
## 95% CI: [ 0.126 ,  0.318 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o4-mini
## ----------------------------------------
## Proportions:  0.63  vs  0.525
## Difference:  0.105
## Chi-squared:  4.797
## Degrees of freedom:  1
## P-value:  0.0285
## 95% CI: [ 0.008 ,  0.202 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.63  vs  0.509
## Difference:  0.121
## Chi-squared:  6.402
## Degrees of freedom:  1
## P-value:  0.0114
## 95% CI: [ 0.024 ,  0.218 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.63  vs  0.343
## Difference:  0.288
## Chi-squared:  37.486
## Degrees of freedom:  1
## P-value:  0.0000000009206
## 95% CI: [ 0.195 ,  0.381 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.63  vs  0.342
## Difference:  0.288
## Chi-squared:  19.096
## Degrees of freedom:  1
## P-value:  0.00001243
## 95% CI: [ 0.157 ,  0.419 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Sonnet-4
## ----------------------------------------
## Proportions:  0.63  vs  0.455
## Difference:  0.175
## Chi-squared:  13.659
## Degrees of freedom:  1
## P-value:  0.0002192
## 95% CI: [ 0.078 ,  0.272 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Opus-4.1
## ----------------------------------------
## Proportions:  0.63  vs  0.741
## Difference:  -0.111
## Chi-squared:  2.601
## Degrees of freedom:  1
## P-value:  0.1068
## 95% CI: [ -0.233 ,  0.011 ]
## Significant:  NO
##
##  Humans vs GPT-5
## ----------------------------------------
## Proportions:  0.63  vs  0.766
## Difference:  -0.136
## Chi-squared:  8.354
## Degrees of freedom:  1
## P-value:  0.003847
## 95% CI: [ -0.22 ,  -0.052 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o3-GPT-Image
## ----------------------------------------
## Proportions:  0.611  vs  0.56
## Difference:  0.05
## Chi-squared:  0.869
## Degrees of freedom:  1
## P-value:  0.3511
## 95% CI: [ -0.05 ,  0.15 ]
## Significant:  NO
##
##  o3 vs o3-Pro
## ----------------------------------------
## Proportions:  0.611  vs  0.772
## Difference:  -0.161
## Chi-squared:  10.208
## Degrees of freedom:  1
## P-value:  0.001398
## 95% CI: [ -0.261 ,  -0.062 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs GPT-4.1
## ----------------------------------------
## Proportions:  0.611  vs  0.47
## Difference:  0.14
## Chi-squared:  5.198
## Degrees of freedom:  1
## P-value:  0.02261
## 95% CI: [ 0.019 ,  0.262 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.611  vs  0.342
## Difference:  0.269
## Chi-squared:  19.762
## Degrees of freedom:  1
## P-value:  0.000008772
## 95% CI: [ 0.151 ,  0.387 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.611  vs  0.408
## Difference:  0.202
## Chi-squared:  11.039
## Degrees of freedom:  1
## P-value:  0.0008922
## 95% CI: [ 0.082 ,  0.322 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o4-mini
## ----------------------------------------
## Proportions:  0.611  vs  0.525
## Difference:  0.085
## Chi-squared:  1.819
## Degrees of freedom:  1
## P-value:  0.1774
## 95% CI: [ -0.036 ,  0.207 ]
## Significant:  NO
##
##  o3 vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.611  vs  0.509
## Difference:  0.101
## Chi-squared:  2.609
## Degrees of freedom:  1
## P-value:  0.1063
## 95% CI: [ -0.02 ,  0.222 ]
## Significant:  NO
##
##  o3 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.611  vs  0.343
## Difference:  0.268
## Chi-squared:  19.636
## Degrees of freedom:  1
## P-value:  0.000009367
## 95% CI: [ 0.15 ,  0.386 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.611  vs  0.342
## Difference:  0.268
## Chi-squared:  11.993
## Degrees of freedom:  1
## P-value:  0.0005341
## 95% CI: [ 0.118 ,  0.419 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Sonnet-4
## ----------------------------------------
## Proportions:  0.611  vs  0.455
## Difference:  0.155
## Chi-squared:  6.383
## Degrees of freedom:  1
## P-value:  0.01152
## 95% CI: [ 0.034 ,  0.276 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Opus-4.1
## ----------------------------------------
## Proportions:  0.611  vs  0.741
## Difference:  -0.131
## Chi-squared:  2.798
## Degrees of freedom:  1
## P-value:  0.09441
## 95% CI: [ -0.273 ,  0.012 ]
## Significant:  NO
##
##  o3 vs GPT-5
## ----------------------------------------
## Proportions:  0.611  vs  0.766
## Difference:  -0.156
## Chi-squared:  7.237
## Degrees of freedom:  1
## P-value:  0.007141
## 95% CI: [ -0.267 ,  -0.045 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs o3-Pro
## ----------------------------------------
## Proportions:  0.56  vs  0.772
## Difference:  -0.211
## Chi-squared:  19.304
## Degrees of freedom:  1
## P-value:  0.00001115
## 95% CI: [ -0.304 ,  -0.119 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs GPT-4.1
## ----------------------------------------
## Proportions:  0.56  vs  0.47
## Difference:  0.09
## Chi-squared:  2.268
## Degrees of freedom:  1
## P-value:  0.132
## 95% CI: [ -0.025 ,  0.206 ]
## Significant:  NO
```

```
##
##  o3-GPT-Image vs GPT-4.1-GPT-Image
## ---------------------------------------
## Proportions:  0.56  vs  0.342
## Difference:  0.219
## Chi-squared:  14.45
## Degrees of freedom:  1
## P-value:  0.000144
## 95% CI: [ 0.107 ,  0.33 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs ChatGPT-4o
## ---------------------------------------
## Proportions:  0.56  vs  0.408
## Difference:  0.152
## Chi-squared:  6.816
## Degrees of freedom:  1
## P-value:  0.009032
## 95% CI: [ 0.038 ,  0.266 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs o4-mini
## ---------------------------------------
## Proportions:  0.56  vs  0.525
## Difference:  0.035
## Chi-squared:  0.272
## Degrees of freedom:  1
## P-value:  0.6019
## 95% CI: [ -0.08 ,  0.151 ]
## Significant:  NO
##
##  o3-GPT-Image vs Gemini-2.5-Pro
## ---------------------------------------
## Proportions:  0.56  vs  0.509
## Difference:  0.051
## Chi-squared:  0.645
## Degrees of freedom:  1
## P-value:  0.422
## 95% CI: [ -0.065 ,  0.167 ]
## Significant:  NO
##
##  o3-GPT-Image vs Gemini-2.0-Flash
## ---------------------------------------
## Proportions:  0.56  vs  0.343
## Difference:  0.218
## Chi-squared:  14.336
## Degrees of freedom:  1
## P-value:  0.0001529
## 95% CI: [ 0.106 ,  0.33 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Gemini-2.0-Flash-Images
## ---------------------------------------
## Proportions:  0.56  vs  0.342
## Difference:  0.218
## Chi-squared:  8.286
## Degrees of freedom:  1
## P-value:  0.003994
## 95% CI: [ 0.072 ,  0.364 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Sonnet-4
## ----------------------------------------
## Proportions:  0.56  vs  0.455
## Difference:  0.105
## Chi-squared:  3.123
## Degrees of freedom:  1
## P-value:  0.0772
## 95% CI: [ -0.01 ,  0.22 ]
## Significant:  NO
##
##  o3-GPT-Image vs Opus-4.1
## ----------------------------------------
## Proportions:  0.56  vs  0.741
## Difference:  -0.181
## Chi-squared:  5.787
## Degrees of freedom:  1
## P-value:  0.01614
## 95% CI: [ -0.319 ,  -0.043 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs GPT-5
## ----------------------------------------
## Proportions:  0.56  vs  0.766
## Difference:  -0.206
## Chi-squared:  13.665
## Degrees of freedom:  1
## P-value:  0.0002185
## 95% CI: [ -0.311 ,  -0.101 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-4.1
## ----------------------------------------
## Proportions:  0.772  vs  0.47
## Difference:  0.302
## Chi-squared:  27.526
## Degrees of freedom:  1
## P-value:  0.000000155
## 95% CI: [ 0.186 ,  0.417 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.772  vs  0.342
## Difference:  0.43
## Chi-squared:  53.691
## Degrees of freedom:  1
## P-value:  0.0000000000002346
## 95% CI: [ 0.318 ,  0.542 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.772  vs  0.408
## Difference:  0.364
## Chi-squared:  39.118
## Degrees of freedom:  1
## P-value:  0.000000000399
## 95% CI: [ 0.249 ,  0.478 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o4-mini
## ----------------------------------------
## Proportions:  0.772  vs  0.525
## Difference:  0.247
## Chi-squared:  18.799
## Degrees of freedom:  1
## P-value:  0.00001452
## 95% CI: [ 0.131 ,  0.362 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.772  vs  0.509
## Difference:  0.262
## Chi-squared:  21.137
## Degrees of freedom:  1
## P-value:  0.000004277
## 95% CI: [ 0.147 ,  0.378 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.772  vs  0.343
## Difference:  0.429
## Chi-squared:  53.494
## Degrees of freedom:  1
## P-value:  0.0000000000002593
## 95% CI: [ 0.318 ,  0.541 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.772  vs  0.342
## Difference:  0.429
## Chi-squared:  35.311
## Degrees of freedom:  1
## P-value:  0.00000000281
## 95% CI: [ 0.283 ,  0.575 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Sonnet-4
## ----------------------------------------
## Proportions:  0.772  vs  0.455
## Difference:  0.316
## Chi-squared:  30.096
## Degrees of freedom:  1
## P-value:  0.00000004112
## 95% CI: [ 0.201 ,  0.431 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Opus-4.1
## ----------------------------------------
## Proportions:  0.772  vs  0.741
## Difference:  0.031
## Chi-squared:  0.095
## Degrees of freedom:  1
## P-value:  0.7581
## 95% CI: [ -0.107 ,  0.168 ]
## Significant:  NO
##
##  o3-Pro vs GPT-5
## ----------------------------------------
## Proportions:  0.772  vs  0.766
## Difference:  0.005
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.097 ,  0.108 ]
## Significant:  NO
##
##  GPT-4.1 vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.47  vs  0.342
## Difference:  0.128
## Chi-squared:  3.587
## Degrees of freedom:  1
## P-value:  0.05825
## 95% CI: [ -0.003 ,  0.26 ]
## Significant:  NO
##
##  GPT-4.1 vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.47  vs  0.408
## Difference:  0.062
## Chi-squared:  0.699
## Degrees of freedom:  1
## P-value:  0.4032
## 95% CI: [ -0.072 ,  0.196 ]
## Significant:  NO
##
##  GPT-4.1 vs o4-mini
## ----------------------------------------
## Proportions:  0.47  vs  0.525
## Difference:  -0.055
## Chi-squared:  0.523
## Degrees of freedom:  1
## P-value:  0.4696
## 95% CI: [ -0.19 ,  0.08 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.47  vs  0.509
## Difference:  -0.039
## Chi-squared:  0.23
## Degrees of freedom:  1
## P-value:  0.6312
## 95% CI: [ -0.174 ,  0.095 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.47  vs  0.343
## Difference:  0.128
## Chi-squared:  3.536
## Degrees of freedom:  1
## P-value:  0.06006
## 95% CI: [ -0.004 ,  0.259 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.47  vs  0.342
## Difference:  0.128
## Chi-squared:  2.171
## Degrees of freedom:  1
## P-value:  0.1406
## 95% CI: [ -0.034 ,  0.29 ]
## Significant:  NO
##
##  GPT-4.1 vs Sonnet-4
## ----------------------------------------
## Proportions:  0.47  vs  0.455
## Difference:  0.015
## Chi-squared:  0.01
## Degrees of freedom:  1
## P-value:  0.9222
## 95% CI: [ -0.12 ,  0.149 ]
## Significant:  NO
##
##  GPT-4.1 vs Opus-4.1
## ----------------------------------------
## Proportions:  0.47  vs  0.741
## Difference:  -0.271
## Chi-squared:  10.854
## Degrees of freedom:  1
## P-value:  0.0009857
## 95% CI: [ -0.426 ,  -0.116 ]
## Significant:  YES (p < 0.05)
```

```
##
##  GPT-4.1 vs GPT-5
## ----------------------------------------
## Proportions:  0.47  vs  0.766
## Difference:  -0.296
## Chi-squared:  21.063
## Degrees of freedom:  1
## P-value:  0.000004444
## 95% CI: [ -0.422 ,  -0.171 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.342  vs  0.408
## Difference:  -0.067
## Chi-squared:  0.866
## Degrees of freedom:  1
## P-value:  0.3519
## 95% CI: [ -0.197 ,  0.064 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs o4-mini
## ----------------------------------------
## Proportions:  0.342  vs  0.525
## Difference:  -0.183
## Chi-squared:  7.489
## Degrees of freedom:  1
## P-value:  0.006208
## 95% CI: [ -0.315 ,  -0.052 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.342  vs  0.509
## Difference:  -0.168
## Chi-squared:  6.234
## Degrees of freedom:  1
## P-value:  0.01253
## 95% CI: [ -0.299 ,  -0.036 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.342  vs  0.343
## Difference:  -0.001
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.122 ,  0.12 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.342  vs  0.342
## Difference:  -0.001
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.148 ,  0.147 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Sonnet-4
## ----------------------------------------
## Proportions:  0.342  vs  0.455
## Difference:  -0.114
## Chi-squared:  2.783
## Degrees of freedom:  1
## P-value:  0.09529
## 95% CI: [ -0.245 ,  0.018 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Opus-4.1
## ----------------------------------------
## Proportions:  0.342  vs  0.741
## Difference:  -0.399
## Chi-squared:  24.017
## Degrees of freedom:  1
## P-value:  0.0000009548
## 95% CI: [ -0.552 ,  -0.247 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs GPT-5
## ----------------------------------------
## Proportions:  0.342  vs  0.766
## Difference:  -0.425
## Chi-squared:  42.073
## Degrees of freedom:  1
## P-value:  0.00000000008791
## 95% CI: [ -0.547 ,  -0.303 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs o4-mini
## ----------------------------------------
## Proportions:  0.408  vs  0.525
## Difference:  -0.117
## Chi-squared:  2.841
## Degrees of freedom:  1
## P-value:  0.09188
## 95% CI: [ -0.251 ,  0.017 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.408  vs  0.509
## Difference:  -0.101
## Chi-squared:  2.084
## Degrees of freedom:  1
## P-value:  0.1488
## 95% CI: [ -0.235 ,  0.033 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.408  vs  0.343
## Difference:  0.066
## Chi-squared:  0.841
## Degrees of freedom:  1
## P-value:  0.359
## 95% CI: [ -0.065 ,  0.196 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.408  vs  0.342
## Difference:  0.066
## Chi-squared:  0.481
## Degrees of freedom:  1
## P-value:  0.4881
## 95% CI: [ -0.095 ,  0.227 ]
## Significant:  NO
##
##  ChatGPT-4o vs Sonnet-4
## ----------------------------------------
## Proportions:  0.408  vs  0.455
## Difference:  -0.047
## Chi-squared:  0.371
## Degrees of freedom:  1
## P-value:  0.5427
## 95% CI: [ -0.181 ,  0.086 ]
## Significant:  NO
##
##  ChatGPT-4o vs Opus-4.1
## ----------------------------------------
## Proportions:  0.408  vs  0.741
## Difference:  -0.333
## Chi-squared:  16.453
## Degrees of freedom:  1
## P-value:  0.00004987
## 95% CI: [ -0.487 ,  -0.179 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs GPT-5
## ----------------------------------------
## Proportions:  0.408  vs  0.766
## Difference:  -0.358
## Chi-squared:  30.278
## Degrees of freedom:  1
## P-value:  0.00000003744
## 95% CI: [ -0.482 ,  -0.234 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.525  vs  0.509
## Difference:  0.016
## Chi-squared:  0.013
## Degrees of freedom:  1
## P-value:  0.9092
## 95% CI: [ -0.119 ,  0.15 ]
## Significant:  NO
##
##  o4-mini vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.525  vs  0.343
## Difference:  0.183
## Chi-squared:  7.416
## Degrees of freedom:  1
## P-value:  0.006465
## 95% CI: [ 0.051 ,  0.314 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.525  vs  0.342
## Difference:  0.183
## Chi-squared:  4.663
## Degrees of freedom:  1
## P-value:  0.03083
## 95% CI: [ 0.021 ,  0.345 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Sonnet-4
## ----------------------------------------
## Proportions:  0.525  vs  0.455
## Difference:  0.07
## Chi-squared:  0.902
## Degrees of freedom:  1
## P-value:  0.3422
## 95% CI: [ -0.065 ,  0.204 ]
## Significant:  NO
##
##  o4-mini vs Opus-4.1
## ----------------------------------------
## Proportions:  0.525  vs  0.741
## Difference:  -0.216
## Chi-squared:  6.888
## Degrees of freedom:  1
## P-value:  0.008676
## 95% CI: [ -0.371 ,  -0.061 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs GPT-5
## ----------------------------------------
## Proportions:  0.525  vs  0.766
## Difference:  -0.241
## Chi-squared:  14.219
## Degrees of freedom:  1
## P-value:  0.0001627
## 95% CI: [ -0.367 ,  -0.116 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5-Pro vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.509  vs  0.343
## Difference:  0.167
## Chi-squared:  6.168
## Degrees of freedom:  1
## P-value:  0.01301
## 95% CI: [ 0.035 ,  0.299 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5-Pro vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.509  vs  0.342
## Difference:  0.167
## Chi-squared:  3.856
## Degrees of freedom:  1
## P-value:  0.04957
## 95% CI: [ 0.005 ,  0.329 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5-Pro vs Sonnet-4
## ----------------------------------------
## Proportions:  0.509  vs  0.455
## Difference:  0.054
## Chi-squared:  0.5
## Degrees of freedom:  1
## P-value:  0.4796
## 95% CI: [ -0.081 ,  0.189 ]
## Significant:  NO
##
##  Gemini-2.5-Pro vs Opus-4.1
## ----------------------------------------
## Proportions:  0.509  vs  0.741
## Difference:  -0.232
## Chi-squared:  7.928
## Degrees of freedom:  1
## P-value:  0.004867
## 95% CI: [ -0.387 ,  -0.077 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5-Pro vs GPT-5
## ----------------------------------------
## Proportions:  0.509  vs  0.766
## Difference:  -0.257
## Chi-squared:  16.044
## Degrees of freedom:  1
## P-value:  0.00006187
## 95% CI: [ -0.382 ,  -0.131 ]
## Significant:  YES (p < 0.05)
```

```
##
##  Gemini-2.0-Flash vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.343  vs  0.342
## Difference:  0
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.147 ,  0.147 ]
## Significant:  NO
##
##  Gemini-2.0-Flash vs Sonnet-4
## ----------------------------------------
## Proportions:  0.343  vs  0.455
## Difference:  -0.113
## Chi-squared:  2.738
## Degrees of freedom:  1
## P-value:  0.098
## 95% CI: [ -0.244 ,  0.018 ]
## Significant:  NO
##
##  Gemini-2.0-Flash vs Opus-4.1
## ----------------------------------------
## Proportions:  0.343  vs  0.741
## Difference:  -0.399
## Chi-squared:  23.911
## Degrees of freedom:  1
## P-value:  0.000001009
## 95% CI: [ -0.551 ,  -0.246 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs GPT-5
## ----------------------------------------
## Proportions:  0.343  vs  0.766
## Difference:  -0.424
## Chi-squared:  41.913
## Degrees of freedom:  1
## P-value:  0.00000000009544
## 95% CI: [ -0.546 ,  -0.302 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-Images vs Sonnet-4
## ----------------------------------------
## Proportions:  0.342  vs  0.455
## Difference:  -0.113
## Chi-squared:  1.665
## Degrees of freedom:  1
## P-value:  0.1969
## 95% CI: [ -0.275 ,  0.049 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-Images vs Opus-4.1
## ----------------------------------------
## Proportions:  0.342  vs  0.741
## Difference:  -0.399
## Chi-squared:  17.647
## Degrees of freedom:  1
## P-value:  0.0000266
## 95% CI: [ -0.579 ,  -0.219 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-Images vs GPT-5
## ----------------------------------------
## Proportions:  0.342  vs  0.766
## Difference:  -0.424
## Chi-squared:  28.89
## Degrees of freedom:  1
## P-value:  0.0000000766
## 95% CI: [ -0.578 ,  -0.27 ]
## Significant:  YES (p < 0.05)
##
##  Sonnet-4 vs Opus-4.1
## ----------------------------------------
## Proportions:  0.455  vs  0.741
## Difference:  -0.286
## Chi-squared:  12.064
## Degrees of freedom:  1
## P-value:  0.0005141
## 95% CI: [ -0.44 ,  -0.131 ]
## Significant:  YES (p < 0.05)
##
##  Sonnet-4 vs GPT-5
## ----------------------------------------
## Proportions:  0.455  vs  0.766
## Difference:  -0.311
## Chi-squared:  23.094
## Degrees of freedom:  1
## P-value:  0.000001543
## 95% CI: [ -0.436 ,  -0.186 ]
## Significant:  YES (p < 0.05)
##
##  Opus-4.1 vs GPT-5
## ----------------------------------------
## Proportions:  0.741  vs  0.766
## Difference:  -0.025
## Chi-squared:  0.035
## Degrees of freedom:  1
## P-value:  0.852
## 95% CI: [ -0.172 ,  0.122 ]
## 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  signifi
## ------------  ---------------------------------------------  -------  ------------  ---------  -------
## X-squared     Humans vs o3                                     0.020     0.1891561     0.6636  FALSE
## X-squared1    Humans vs o3-GPT-Image                           0.070     4.0094812     0.0452  TRUE
## X-squared2    Humans vs o3-Pro                                -0.141    13.4669053     0.0002  TRUE
## X-squared3    Humans vs GPT-4.1                                0.160    11.4260726     0.0007  TRUE
## X-squared4    Humans vs GPT-4.1-GPT-Image                      0.289    37.7050952     0.0000  TRUE
## X-squared5    Humans vs ChatGPT-4o                             0.222    22.2173252     0.0000  TRUE
## X-squared6    Humans vs o4-mini                                0.105     4.7972861     0.0285  TRUE
## X-squared7    Humans vs Gemini-2.5-Pro                         0.121     6.4021831     0.0114  TRUE
## X-squared8    Humans vs Gemini-2.0-Flash                       0.288    37.4862876     0.0000  TRUE
## X-squared9    Humans vs Gemini-2.0-Flash-Images                0.288    19.0963849     0.0000  TRUE
## X-squared10   Humans vs Sonnet-4                               0.175    13.6589132     0.0002  TRUE
## X-squared11   Humans vs Opus-4.1                              -0.111     2.6009376     0.1068  FALSE
## X-squared12   Humans vs GPT-5                                 -0.136     8.3544955     0.0038  TRUE
## X-squared13   o3 vs o3-GPT-Image                               0.050     0.8693549     0.3511  FALSE
## X-squared14   o3 vs o3-Pro                                    -0.161    10.2079814     0.0014  TRUE
## X-squared15   o3 vs GPT-4.1                                    0.140     5.1982122     0.0226  TRUE
## X-squared16   o3 vs GPT-4.1-GPT-Image                          0.269    19.7618003     0.0000  TRUE
## X-squared17   o3 vs ChatGPT-4o                                 0.202    11.0389794     0.0009  TRUE
## X-squared18   o3 vs o4-mini                                    0.085     1.8190399     0.1774  FALSE
## X-squared19   o3 vs Gemini-2.5-Pro                             0.101     2.6088228     0.1063  FALSE
## X-squared20   o3 vs Gemini-2.0-Flash                           0.268    19.6364283     0.0000  TRUE
## X-squared21   o3 vs Gemini-2.0-Flash-Images                    0.268    11.9927811     0.0005  TRUE
## X-squared22   o3 vs Sonnet-4                                   0.155     6.3833257     0.0115  TRUE
## X-squared23   o3 vs Opus-4.1                                  -0.131     2.7975028     0.0944  FALSE
## X-squared24   o3 vs GPT-5                                     -0.156     7.2371887     0.0071  TRUE
## X-squared25   o3-GPT-Image vs o3-Pro                          -0.211    19.3040615     0.0000  TRUE
## X-squared26   o3-GPT-Image vs GPT-4.1                          0.090     2.2682502     0.1320  FALSE
## X-squared27   o3-GPT-Image vs GPT-4.1-GPT-Image                0.219    14.4496552     0.0001  TRUE
## X-squared28   o3-GPT-Image vs ChatGPT-4o                       0.152     6.8164163     0.0090  TRUE
## X-squared29   o3-GPT-Image vs o4-mini                          0.035     0.2722063     0.6019  FALSE
## X-squared30   o3-GPT-Image vs Gemini-2.5-Pro                   0.051     0.6448736     0.4220  FALSE
## X-squared31   o3-GPT-Image vs Gemini-2.0-Flash                 0.218    14.3361143     0.0002  TRUE
## X-squared32   o3-GPT-Image vs Gemini-2.0-Flash-Images          0.218     8.2863655     0.0040  TRUE
## X-squared33   o3-GPT-Image vs Sonnet-4                         0.105     3.1229042     0.0772  FALSE
## X-squared34   o3-GPT-Image vs Opus-4.1                        -0.181     5.7870517     0.0161  TRUE
## X-squared35   o3-GPT-Image vs GPT-5                           -0.206    13.6649251     0.0002  TRUE
## X-squared36   o3-Pro vs GPT-4.1                                0.302    27.5259711     0.0000  TRUE
## X-squared37   o3-Pro vs GPT-4.1-GPT-Image                      0.430    53.6909027     0.0000  TRUE
## X-squared38   o3-Pro vs ChatGPT-4o                             0.364    39.1177375     0.0000  TRUE
## X-squared39   o3-Pro vs o4-mini                                0.247    18.7992418     0.0000  TRUE
## X-squared40   o3-Pro vs Gemini-2.5-Pro                         0.262    21.1366991     0.0000  TRUE
## X-squared41   o3-Pro vs Gemini-2.0-Flash                       0.429    53.4944490     0.0000  TRUE
## X-squared42   o3-Pro vs Gemini-2.0-Flash-Images                0.429    35.3111352     0.0000  TRUE
## X-squared43   o3-Pro vs Sonnet-4                               0.316    30.0957775     0.0000  TRUE
## X-squared44   o3-Pro vs Opus-4.1                               0.031     0.0948295     0.7581  FALSE
## X-squared45   o3-Pro vs GPT-5                                 0.005     0.0000000    1.0000  FALSE
## X-squared46   GPT-4.1 vs GPT-4.1-GPT-Image                    0.128     3.5865624    0.0582  FALSE
## X-squared47   GPT-4.1 vs ChatGPT-4o                           0.062     0.6986238    0.4032  FALSE
## X-squared48   GPT-4.1 vs o4-mini                             -0.055     0.5229013    0.4696  FALSE
## X-squared49   GPT-4.1 vs Gemini-2.5-Pro                      -0.039     0.2304718    0.6312  FALSE
## X-squared50   GPT-4.1 vs Gemini-2.0-Flash                     0.128     3.5357988    0.0601  FALSE
## X-squared51   GPT-4.1 vs Gemini-2.0-Flash-Images              0.128     2.1711852    0.1406  FALSE
## X-squared52   GPT-4.1 vs Sonnet-4                             0.015     0.0095474    0.9222  FALSE
## X-squared53   GPT-4.1 vs Opus-4.1                            -0.271    10.8542730    0.0010  TRUE
## X-squared54   GPT-4.1 vs GPT-5                               -0.296    21.0630216    0.0000  TRUE
## X-squared55   GPT-4.1-GPT-Image vs ChatGPT-4o                -0.067     0.8664502    0.3519  FALSE
## X-squared56   GPT-4.1-GPT-Image vs o4-mini                   -0.183     7.4888306    0.0062  TRUE
## X-squared57   GPT-4.1-GPT-Image vs Gemini-2.5-Pro            -0.168     6.2344269    0.0125  TRUE
## X-squared58   GPT-4.1-GPT-Image vs Gemini-2.0-Flash          -0.001     0.0000000    1.0000  FALSE
## X-squared59   GPT-4.1-GPT-Image vs Gemini-2.0-Flash-Images   -0.001     0.0000000    1.0000  FALSE
## X-squared60   GPT-4.1-GPT-Image vs Sonnet-4                  -0.114     2.7825972    0.0953  FALSE
## X-squared61   GPT-4.1-GPT-Image vs Opus-4.1                  -0.399    24.0170961    0.0000  TRUE
## X-squared62   GPT-4.1-GPT-Image vs GPT-5                     -0.425    42.0733617    0.0000  TRUE
## X-squared63   ChatGPT-4o vs o4-mini                          -0.117     2.8412057    0.0919  FALSE
## X-squared64   ChatGPT-4o vs Gemini-2.5-Pro                   -0.101     2.0839971    0.1488  FALSE
## X-squared65   ChatGPT-4o vs Gemini-2.0-Flash                  0.066     0.8414311    0.3590  FALSE
## X-squared66   ChatGPT-4o vs Gemini-2.0-Flash-Images           0.066     0.4806813    0.4881  FALSE
## X-squared67   ChatGPT-4o vs Sonnet-4                         -0.047     0.3705979    0.5427  FALSE
## X-squared68   ChatGPT-4o vs Opus-4.1                         -0.333    16.4528721    0.0000  TRUE
## X-squared69   ChatGPT-4o vs GPT-5                            -0.358    30.2778543    0.0000  TRUE
## X-squared70   o4-mini vs Gemini-2.5-Pro                       0.016     0.0130201    0.9092  FALSE
## X-squared71   o4-mini vs Gemini-2.0-Flash                     0.183     7.4160347    0.0065  TRUE
## X-squared72   o4-mini vs Gemini-2.0-Flash-Images              0.183     4.6625568    0.0308  TRUE
## X-squared73   o4-mini vs Sonnet-4                             0.070     0.9021768    0.3422  FALSE
## X-squared74   o4-mini vs Opus-4.1                            -0.216     6.8883699    0.0087  TRUE
## X-squared75   o4-mini vs GPT-5                               -0.241    14.2190034    0.0002  TRUE
## X-squared76   Gemini-2.5-Pro vs Gemini-2.0-Flash              0.167     6.1678391    0.0130  TRUE
## X-squared77   Gemini-2.5-Pro vs Gemini-2.0-Flash-Images       0.167     3.8559573    0.0496  TRUE
## X-squared78   Gemini-2.5-Pro vs Sonnet-4                      0.054     0.4997782    0.4796  FALSE
## X-squared79   Gemini-2.5-Pro vs Opus-4.1                     -0.232     7.9282614    0.0049  TRUE
## X-squared80   Gemini-2.5-Pro vs GPT-5                        -0.257    16.0443888    0.0001  TRUE
## X-squared81   Gemini-2.0-Flash vs Gemini-2.0-Flash-Images     0.000     0.0000000    1.0000  FALSE
## X-squared82   Gemini-2.0-Flash vs Sonnet-4                   -0.113     2.7378221    0.0980  FALSE
## X-squared83   Gemini-2.0-Flash vs Opus-4.1                   -0.399    23.9111061    0.0000  TRUE
## X-squared84   Gemini-2.0-Flash vs GPT-5                      -0.424    41.9127251    0.0000  TRUE
## X-squared85   Gemini-2.0-Flash-Images vs Sonnet-4            -0.113     1.6654778    0.1969  FALSE
## X-squared86   Gemini-2.0-Flash-Images vs Opus-4.1            -0.399    17.6467601    0.0000  TRUE
## X-squared87   Gemini-2.0-Flash-Images vs GPT-5               -0.424    28.8900877    0.0000  TRUE
## X-squared88   Sonnet-4 vs Opus-4.1                           -0.286    12.0637483    0.0005  TRUE
## X-squared89   Sonnet-4 vs GPT-5                              -0.311    23.0935495    0.0000  TRUE
## X-squared90   Opus-4.1 vs GPT-5                              -0.025     0.0348205    0.8520  FALSE
```

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:  0.526  vs  0.649 
## Difference:  -0.123 
## Chi-squared:  38.861 
## Degrees of freedom:  1 
## P-value:  0.0000000004552 
## 95% CI: [ -0.161 ,  -0.086 ]
## Significant:  YES (p < 0.05) 
## 
##  Humans vs o3-GPT-Image 
## ---------------------------------------- 
## Proportions:  0.526  vs  0.552 
## Difference:  -0.026 
## Chi-squared:  2.115 
## Degrees of freedom:  1 
## P-value:  0.1458 
## 95% CI: [ -0.06 ,  0.009 ]
## Significant:  NO 
## 
##  Humans vs o3-Pro 
## ---------------------------------------- 
## Proportions:  0.526  vs  0.64 
## Difference:  -0.114 
## Chi-squared:  33.112 
## Degrees of freedom:  1 
## P-value:  0.000000008702 
## 95% CI: [ -0.152 ,  -0.076 ]
## Significant:  YES (p < 0.05) 
## 
##  Humans vs GPT-4.1 
## ---------------------------------------- 
## Proportions:  0.526  vs  0.413 
## Difference:  0.112
## Chi-squared:  22.059
## Degrees of freedom:  1
## P-value:  0.000002644
## 95% CI: [ 0.066 ,  0.159 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.526  vs  0.393
## Difference:  0.133
## Chi-squared:  30.716
## Degrees of freedom:  1
## P-value:  0.00000002987
## 95% CI: [ 0.086 ,  0.179 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.526  vs  0.422
## Difference:  0.103
## Chi-squared:  18.644
## Degrees of freedom:  1
## P-value:  0.00001575
## 95% CI: [ 0.056 ,  0.151 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o4-mini
## ----------------------------------------
## Proportions:  0.526  vs  0.532
## Difference:  -0.006
## Chi-squared:  0.035
## Degrees of freedom:  1
## P-value:  0.8507
## 95% CI: [ -0.053 ,  0.042 ]
## Significant:  NO
##
##  Humans vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.526  vs  0.45
## Difference:  0.076
## Chi-squared:  9.986
## Degrees of freedom:  1
## P-value:  0.001577
## 95% CI: [ 0.029 ,  0.123 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.526  vs  0.389
## Difference:  0.137
## Chi-squared:  32.638
## Degrees of freedom:  1
## P-value:  0.0000000111
## 95% CI: [ 0.09 ,  0.183 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.526  vs  0.328
## Difference:  0.198
## Chi-squared:  35.28
## Degrees of freedom:  1
## P-value:  0.000000002855
## 95% CI: [ 0.135 ,  0.26 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Sonnet-4
## ----------------------------------------
## Proportions:  0.526  vs  0.407
## Difference:  0.119
## Chi-squared:  24.575
## Degrees of freedom:  1
## P-value:  0.0000007149
## 95% CI: [ 0.072 ,  0.166 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Opus-4.1
## ----------------------------------------
## Proportions:  0.526  vs  0.476
## Difference:  0.049
## Chi-squared:  2.068
## Degrees of freedom:  1
## P-value:  0.1504
## 95% CI: [ -0.017 ,  0.116 ]
## Significant:  NO
##
##  Humans vs GPT-5
## ----------------------------------------
## Proportions:  0.526  vs  0.646
## Difference:  -0.12
## Chi-squared:  25.084
## Degrees of freedom:  1
## P-value:  0.0000005489
## 95% CI: [ -0.165 ,  -0.074 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o3-GPT-Image
## ----------------------------------------
## Proportions:  0.649  vs  0.552
## Difference:  0.098
## Chi-squared:  15.818
## Degrees of freedom:  1
## P-value:  0.00006973
## 95% CI: [ 0.049 ,  0.146 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o3-Pro
## ----------------------------------------
## Proportions:  0.649  vs  0.64
## Difference:  0.009
## Chi-squared:  0.101
## Degrees of freedom:  1
## P-value:  0.7504
## 95% CI: [ -0.041 ,  0.06 ]
## Significant:  NO
##
##  o3 vs GPT-4.1
## ----------------------------------------
## Proportions:  0.649  vs  0.413
## Difference:  0.236
## Chi-squared:  63.91
## Degrees of freedom:  1
## P-value:  0.000000000000001303
## 95% CI: [ 0.178 ,  0.294 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.649  vs  0.393
## Difference:  0.256
## Chi-squared:  75.081
## Degrees of freedom:  1
## P-value:  0.000000000000000004518
## 95% CI: [ 0.198 ,  0.314 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.649  vs  0.422
## Difference:  0.227
## Chi-squared:  59.196
## Degrees of freedom:  1
## P-value:  0.00000000000001428
## 95% CI: [ 0.169 ,  0.285 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o4-mini
## ----------------------------------------
## Proportions:  0.649  vs  0.532
## Difference:  0.118
## Chi-squared:  16.191
## Degrees of freedom:  1
## P-value:  0.00005726
## 95% CI: [ 0.059 ,  0.176 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.649  vs  0.45
## Difference:  0.199
## Chi-squared:  45.897
## Degrees of freedom:  1
## P-value:  0.00000000001246
## 95% CI: [ 0.141 ,  0.258 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.649  vs  0.389
## Difference:  0.26
## Chi-squared:  77.448
## Degrees of freedom:  1
## P-value:  0.000000000000000001363
## 95% CI: [ 0.202 ,  0.318 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.649  vs  0.328
## Difference:  0.321
## Chi-squared:  74.296
## Degrees of freedom:  1
## P-value:  0.000000000000000006724
## 95% CI: [ 0.249 ,  0.393 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Sonnet-4
## ----------------------------------------
## Proportions:  0.649  vs  0.407
## Difference:  0.242
## Chi-squared:  67.256
## Degrees of freedom:  1
## P-value:  0.0000000000000002384
## 95% CI: [ 0.184 ,  0.3 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Opus-4.1
## ----------------------------------------
## Proportions:  0.649  vs  0.476
## Difference:  0.173
## Chi-squared:  21.801
## Degrees of freedom:  1
## P-value:  0.000003025
## 95% CI: [ 0.098 ,  0.248 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs GPT-5
## ----------------------------------------
## Proportions:  0.649  vs  0.646
## Difference:  0.004
## Chi-squared:  0.005
## Degrees of freedom:  1
## P-value:  0.9437
## 95% CI: [ -0.053 ,  0.061 ]
## Significant:  NO
```

```
##
##  o3-GPT-Image vs o3-Pro
## ----------------------------------------
## Proportions:  0.552  vs  0.64
## Difference:  -0.088
## Chi-squared:  12.838
## Degrees of freedom:  1
## P-value:  0.0003397
## 95% CI: [ -0.136 ,  -0.04 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs GPT-4.1
## ----------------------------------------
## Proportions:  0.552  vs  0.413
## Difference:  0.138
## Chi-squared:  23.943
## Degrees of freedom:  1
## P-value:  0.0000009922
## 95% CI: [ 0.083 ,  0.194 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.552  vs  0.393
## Difference:  0.158
## Chi-squared:  31.477
## Degrees of freedom:  1
## P-value:  0.00000002018
## 95% CI: [ 0.103 ,  0.214 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.552  vs  0.422
## Difference:  0.129
## Chi-squared:  20.904
## Degrees of freedom:  1
## P-value:  0.000004829
## 95% CI: [ 0.074 ,  0.185 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs o4-mini
## ----------------------------------------
## Proportions:  0.552  vs  0.532
## Difference:  0.02
## Chi-squared:  0.451
## Degrees of freedom:  1
## P-value:  0.5017
## 95% CI: [ -0.036 ,  0.076 ]
## Significant:  NO
##
##  o3-GPT-Image vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.552  vs  0.45
## Difference:  0.102
## Chi-squared:  12.897
## Degrees of freedom:  1
## P-value:  0.0003291
## 95% CI: [ 0.046 ,  0.158 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.552  vs  0.389
## Difference:  0.162
## Chi-squared:  33.126
## Degrees of freedom:  1
## P-value:  0.000000008639
## 95% CI: [ 0.107 ,  0.218 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.552  vs  0.328
## Difference:  0.223
## Chi-squared:  37.45
## Degrees of freedom:  1
## P-value:  0.0000000009377
## 95% CI: [ 0.154 ,  0.293 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Sonnet-4
## ----------------------------------------
## Proportions:  0.552  vs  0.407
## Difference:  0.145
## Chi-squared:  26.154
## Degrees of freedom:  1
## P-value:  0.0000003152
## 95% CI: [ 0.089 ,  0.2 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Opus-4.1
## ----------------------------------------
## Proportions:  0.552  vs  0.476
## Difference:  0.075
## Chi-squared:  4.081
## Degrees of freedom:  1
## P-value:  0.04337
## 95% CI: [ 0.002 ,  0.148 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs GPT-5
## ----------------------------------------
## Proportions:  0.552  vs  0.646
## Difference:  -0.094
## Chi-squared:  11.198
## Degrees of freedom:  1
## P-value:  0.0008187
## 95% CI: [ -0.148 ,  -0.039 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-4.1
## ----------------------------------------
## Proportions:  0.64  vs  0.413
## Difference:  0.226
## Chi-squared:  58.734
## Degrees of freedom:  1
## P-value:  0.00000000000001805
## 95% CI: [ 0.168 ,  0.284 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.64  vs  0.393
## Difference:  0.246
## Chi-squared:  69.483
## Degrees of freedom:  1
## P-value:  0.00000000000000007707
## 95% CI: [ 0.189 ,  0.304 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.64  vs  0.422
## Difference:  0.217
## Chi-squared:  54.21
## Degrees of freedom:  1
## P-value:  0.0000000000001801
## 95% CI: [ 0.159 ,  0.276 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o4-mini
## ----------------------------------------
## Proportions:  0.64  vs  0.532
## Difference:  0.108
## Chi-squared:  13.607
## Degrees of freedom:  1
## P-value:  0.0002253
## 95% CI: [ 0.05 ,  0.167 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.64  vs  0.45
## Difference:  0.19
## Chi-squared:  41.502
## Degrees of freedom:  1
## P-value:  0.0000000001177
## 95% CI: [ 0.132 ,  0.248 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.64  vs  0.389
## Difference:  0.251
## Chi-squared:  71.765
## Degrees of freedom:  1
## P-value:  0.00000000000000002424
## 95% CI: [ 0.193 ,  0.308 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.64  vs  0.328
## Difference:  0.312
## Chi-squared:  69.702
## Degrees of freedom:  1
## P-value:  0.00000000000000006899
## 95% CI: [ 0.24 ,  0.383 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Sonnet-4
## ----------------------------------------
## Proportions:  0.64  vs  0.407
## Difference:  0.233
## Chi-squared:  61.95
## Degrees of freedom:  1
## P-value:  0.000000000000003524
## 95% CI: [ 0.175 ,  0.291 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Opus-4.1
## ----------------------------------------
## Proportions:  0.64  vs  0.476
## Difference:  0.163
## Chi-squared:  19.337
## Degrees of freedom:  1
## P-value:  0.00001096
## 95% CI: [ 0.088 ,  0.238 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-5
## ----------------------------------------
## Proportions:  0.64  vs  0.646
## Difference:  -0.006
## Chi-squared:  0.02
## Degrees of freedom:  1
## P-value:  0.8887
## 95% CI: [ -0.063 ,  0.051 ]
## Significant:  NO
##
##  GPT-4.1 vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.413  vs  0.393
## Difference:  0.02
## Chi-squared:  0.322
## Degrees of freedom:  1
## P-value:  0.5703
## 95% CI: [ -0.044 ,  0.084 ]
## Significant:  NO
##
##  GPT-4.1 vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.413  vs  0.422
## Difference:  -0.009
## Chi-squared:  0.047
## Degrees of freedom:  1
## P-value:  0.8284
## 95% CI: [ -0.073 ,  0.055 ]
## Significant:  NO
##
##  GPT-4.1 vs o4-mini
## ----------------------------------------
## Proportions:  0.413  vs  0.532
## Difference:  -0.118
## Chi-squared:  12.951
## Degrees of freedom:  1
## P-value:  0.0003197
## 95% CI: [ -0.183 ,  -0.053 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.413  vs  0.45
## Difference:  -0.036
## Chi-squared:  1.155
## Degrees of freedom:  1
## P-value:  0.2825
## 95% CI: [ -0.101 ,  0.028 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.413  vs  0.389
## Difference:  0.024
## Chi-squared:  0.485
## Degrees of freedom:  1
## P-value:  0.4863
## 95% CI: [ -0.04 ,  0.088 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.413  vs  0.328
## Difference:  0.085
## Chi-squared:  4.539
## Degrees of freedom:  1
## P-value:  0.03313
## 95% CI: [ 0.008 ,  0.162 ]
## Significant:  YES (p < 0.05)
```

```
##
##  GPT-4.1 vs Sonnet-4
## ----------------------------------------
## Proportions:  0.413  vs  0.407
## Difference:  0.006
## Chi-squared:  0.017
## Degrees of freedom:  1
## P-value:  0.8973
## 95% CI: [ -0.058 ,  0.07 ]
## Significant:  NO
##
##  GPT-4.1 vs Opus-4.1
## ----------------------------------------
## Proportions:  0.413  vs  0.476
## Difference:  -0.063
## Chi-squared:  2.336
## Degrees of freedom:  1
## P-value:  0.1264
## 95% CI: [ -0.143 ,  0.017 ]
## Significant:  NO
##
##  GPT-4.1 vs GPT-5
## ----------------------------------------
## Proportions:  0.413  vs  0.646
## Difference:  -0.232
## Chi-squared:  50.978
## Degrees of freedom:  1
## P-value:  0.0000000000009341
## 95% CI: [ -0.296 ,  -0.169 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.393  vs  0.422
## Difference:  -0.029
## Chi-squared:  0.722
## Degrees of freedom:  1
## P-value:  0.3954
## 95% CI: [ -0.093 ,  0.035 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs o4-mini
## ----------------------------------------
## Proportions:  0.393  vs  0.532
## Difference:  -0.138
## Chi-squared:  17.865
## Degrees of freedom:  1
## P-value:  0.00002371
## 95% CI: [ -0.203 ,  -0.074 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.393  vs  0.45
## Difference:  -0.057
## Chi-squared:  2.915
## Degrees of freedom:  1
## P-value:  0.08778
## 95% CI: [ -0.121 ,  0.008 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.393  vs  0.389
## Difference:  0.004
## Chi-squared:  0.004
## Degrees of freedom:  1
## P-value:  0.9502
## 95% CI: [ -0.06 ,  0.068 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.393  vs  0.328
## Difference:  0.065
## Chi-squared:  2.625
## Degrees of freedom:  1
## P-value:  0.1052
## 95% CI: [ -0.012 ,  0.142 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Sonnet-4
## ----------------------------------------
## Proportions:  0.393  vs  0.407
## Difference:  -0.014
## Chi-squared:  0.139
## Degrees of freedom:  1
## P-value:  0.7092
## 95% CI: [ -0.078 ,  0.05 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Opus-4.1
## ----------------------------------------
## Proportions:  0.393  vs  0.476
## Difference:  -0.083
## Chi-squared:  4.196
## Degrees of freedom:  1
## P-value:  0.04053
## 95% CI: [ -0.163 ,  -0.003 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs GPT-5
## ----------------------------------------
## Proportions:  0.393  vs  0.646
## Difference:  -0.252
## Chi-squared:  60.137
## Degrees of freedom:  1
## P-value:  0.00000000000000885
## 95% CI: [ -0.315 ,  -0.189 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs o4-mini
## ----------------------------------------
## Proportions:  0.422  vs  0.532
## Difference:  -0.109
## Chi-squared:  11.012
## Degrees of freedom:  1
## P-value:  0.0009052
## 95% CI: [ -0.174 ,  -0.044 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.422  vs  0.45
## Difference:  -0.027
## Chi-squared:  0.628
## Degrees of freedom:  1
## P-value:  0.4279
## 95% CI: [ -0.092 ,  0.037 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.422  vs  0.389
## Difference:  0.033
## Chi-squared:  0.957
## Degrees of freedom:  1
## P-value:  0.3279
## 95% CI: [ -0.031 ,  0.097 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.422  vs  0.328
## Difference:  0.094
## Chi-squared:  5.557
## Degrees of freedom:  1
## P-value:  0.01841
## 95% CI: [ 0.017 ,  0.171 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs Sonnet-4
## ----------------------------------------
## Proportions:  0.422  vs  0.407
## Difference:  0.015
## Chi-squared:  0.169
## Degrees of freedom:  1
## P-value:  0.6808
## 95% CI: [ -0.049 ,  0.08 ]
## Significant:  NO
##
##  ChatGPT-4o vs Opus-4.1
## ----------------------------------------
## Proportions:  0.422  vs  0.476
## Difference:  -0.054
## Chi-squared:  1.683
## Degrees of freedom:  1
## P-value:  0.1946
## 95% CI: [ -0.134 ,  0.026 ]
## Significant:  NO
##
##  ChatGPT-4o vs GPT-5
## ----------------------------------------
## Proportions:  0.422  vs  0.646
## Difference:  -0.223
## Chi-squared:  47.127
## Degrees of freedom:  1
## P-value:  0.000000000006653
## 95% CI: [ -0.287 ,  -0.16 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.532  vs  0.45
## Difference:  0.082
## Chi-squared:  6.074
## Degrees of freedom:  1
## P-value:  0.01372
## 95% CI: [ 0.016 ,  0.147 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.532  vs  0.389
## Difference:  0.142
## Chi-squared:  18.957
## Degrees of freedom:  1
## P-value:  0.00001337
## 95% CI: [ 0.078 ,  0.207 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.532  vs  0.328
## Difference:  0.203
## Chi-squared:  25.739
## Degrees of freedom:  1
## P-value:  0.0000003908
## 95% CI: [ 0.126 ,  0.281 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Sonnet-4
## ----------------------------------------
## Proportions:  0.532  vs  0.407
## Difference:  0.124
## Chi-squared:  14.379
## Degrees of freedom:  1
## P-value:  0.0001495
## 95% CI: [ 0.06 ,  0.189 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Opus-4.1
## ----------------------------------------
## Proportions:  0.532  vs  0.476
## Difference:  0.055
## Chi-squared:  1.726
## Degrees of freedom:  1
## P-value:  0.1889
## 95% CI: [ -0.025 ,  0.136 ]
## Significant:  NO
##
##  o4-mini vs GPT-5
## ----------------------------------------
## Proportions:  0.532  vs  0.646
## Difference:  -0.114
## Chi-squared:  12.427
## Degrees of freedom:  1
## P-value:  0.0004231
## 95% CI: [ -0.178 ,  -0.05 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5-Pro vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.45  vs  0.389
## Difference:  0.061
## Chi-squared:  3.369
## Degrees of freedom:  1
## P-value:  0.06642
## 95% CI: [ -0.004 ,  0.125 ]
## Significant:  NO
##
##  Gemini-2.5-Pro vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.45  vs  0.328
## Difference:  0.122
## Chi-squared:  9.277
## Degrees of freedom:  1
## P-value:  0.00232
## 95% CI: [ 0.044 ,  0.199 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5-Pro vs Sonnet-4
## ----------------------------------------
## Proportions:  0.45  vs  0.407
## Difference:  0.043
## Chi-squared:  1.611
## Degrees of freedom:  1
## P-value:  0.2044
## 95% CI: [ -0.022 ,  0.107 ]
## Significant:  NO
```

```
##
##  Gemini-2.5-Pro vs Opus-4.1
## ----------------------------------------
## Proportions:  0.45  vs  0.476
## Difference:  -0.027
## Chi-squared:  0.354
## Degrees of freedom:  1
## P-value:  0.5517
## 95% CI: [ -0.107 ,  0.054 ]
## Significant:  NO
##
##  Gemini-2.5-Pro vs GPT-5
## ----------------------------------------
## Proportions:  0.45  vs  0.646
## Difference:  -0.196
## Chi-squared:  36.309
## Degrees of freedom:  1
## P-value:  0.000000001684
## 95% CI: [ -0.259 ,  -0.132 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.389  vs  0.328
## Difference:  0.061
## Chi-squared:  2.3
## Degrees of freedom:  1
## P-value:  0.1294
## 95% CI: [ -0.016 ,  0.138 ]
## Significant:  NO
##
##  Gemini-2.0-Flash vs Sonnet-4
## ----------------------------------------
## Proportions:  0.389  vs  0.407
## Difference:  -0.018
## Chi-squared:  0.251
## Degrees of freedom:  1
## P-value:  0.6161
## 95% CI: [ -0.082 ,  0.046 ]
## Significant:  NO
##
##  Gemini-2.0-Flash vs Opus-4.1
## ----------------------------------------
## Proportions:  0.389  vs  0.476
## Difference:  -0.087
## Chi-squared:  4.64
## Degrees of freedom:  1
## P-value:  0.03123
## 95% CI: [ -0.167 ,  -0.007 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs GPT-5
## ----------------------------------------
## Proportions:  0.389  vs  0.646
## Difference:  -0.256
## Chi-squared:  62.083
## Degrees of freedom:  1
## P-value:  0.000000000000003293
## 95% CI: [ -0.319 ,  -0.193 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-Images vs Sonnet-4
## ----------------------------------------
## Proportions:  0.328  vs  0.407
## Difference:  -0.079
## Chi-squared:  3.896
## Degrees of freedom:  1
## P-value:  0.0484
## 95% CI: [ -0.156 ,  -0.002 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-Images vs Opus-4.1
## ----------------------------------------
## Proportions:  0.328  vs  0.476
## Difference:  -0.148
## Chi-squared:  10.339
## Degrees of freedom:  1
## P-value:  0.001303
## 95% CI: [ -0.239 ,  -0.057 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-Images vs GPT-5
## ----------------------------------------
## Proportions:  0.328  vs  0.646
## Difference:  -0.317
## Chi-squared:  63.537
## Degrees of freedom:  1
## P-value:  0.000000000000001574
## 95% CI: [ -0.394 ,  -0.241 ]
## Significant:  YES (p < 0.05)
##
##  Sonnet-4 vs Opus-4.1
## ----------------------------------------
## Proportions:  0.407  vs  0.476
## Difference:  -0.069
## Chi-squared:  2.85
## Degrees of freedom:  1
## P-value:  0.0914
## 95% CI: [ -0.149 ,  0.011 ]
## Significant:  NO
##
##  Sonnet-4 vs GPT-5
## ----------------------------------------
## Proportions:  0.407  vs  0.646
## Difference:  -0.238
## Chi-squared:  53.717
## Degrees of freedom:  1
## P-value:  0.0000000000002316
## 95% CI: [ -0.302 ,  -0.175 ]
## Significant:  YES (p < 0.05)
##
##  Opus-4.1 vs GPT-5
## ----------------------------------------
## Proportions:  0.476  vs  0.646
## Difference:  -0.169
## Chi-squared:  18.21
## Degrees of freedom:  1
## P-value:  0.00001978
## 95% CI: [ -0.249 ,  -0.09 ]
## 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")
```

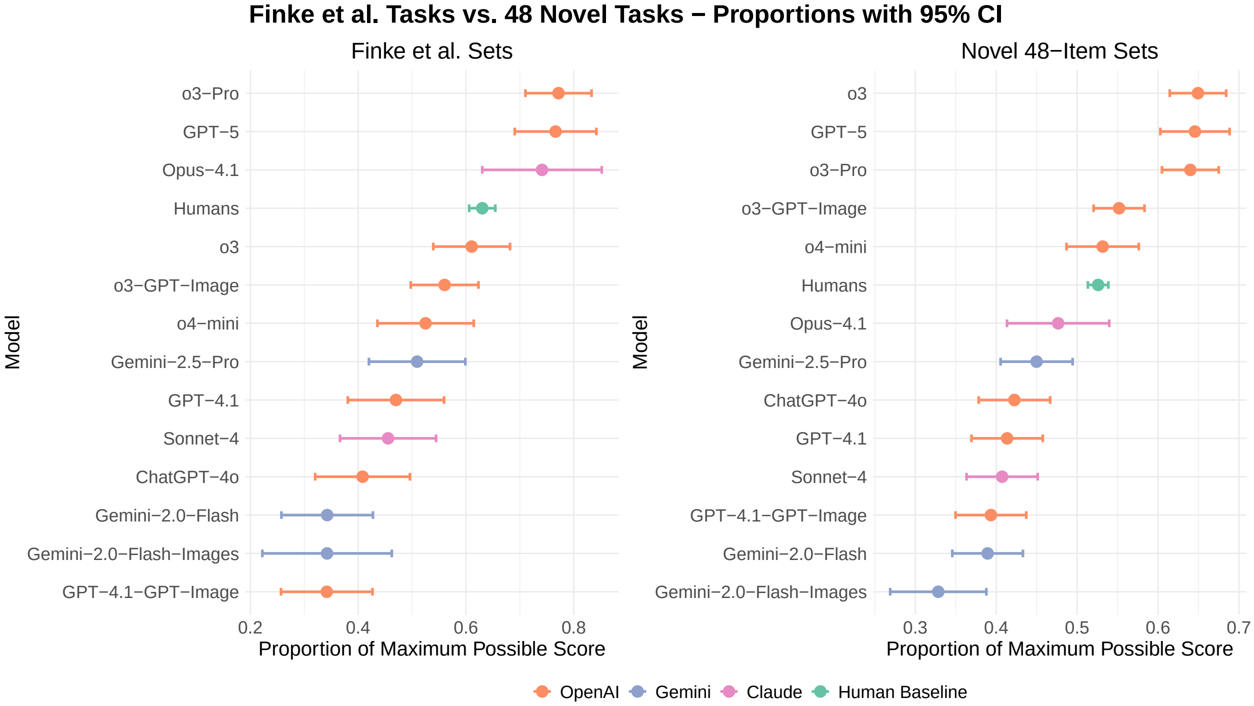
```
##
##
## Summary Table - 48 Novel Tasks:
```

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

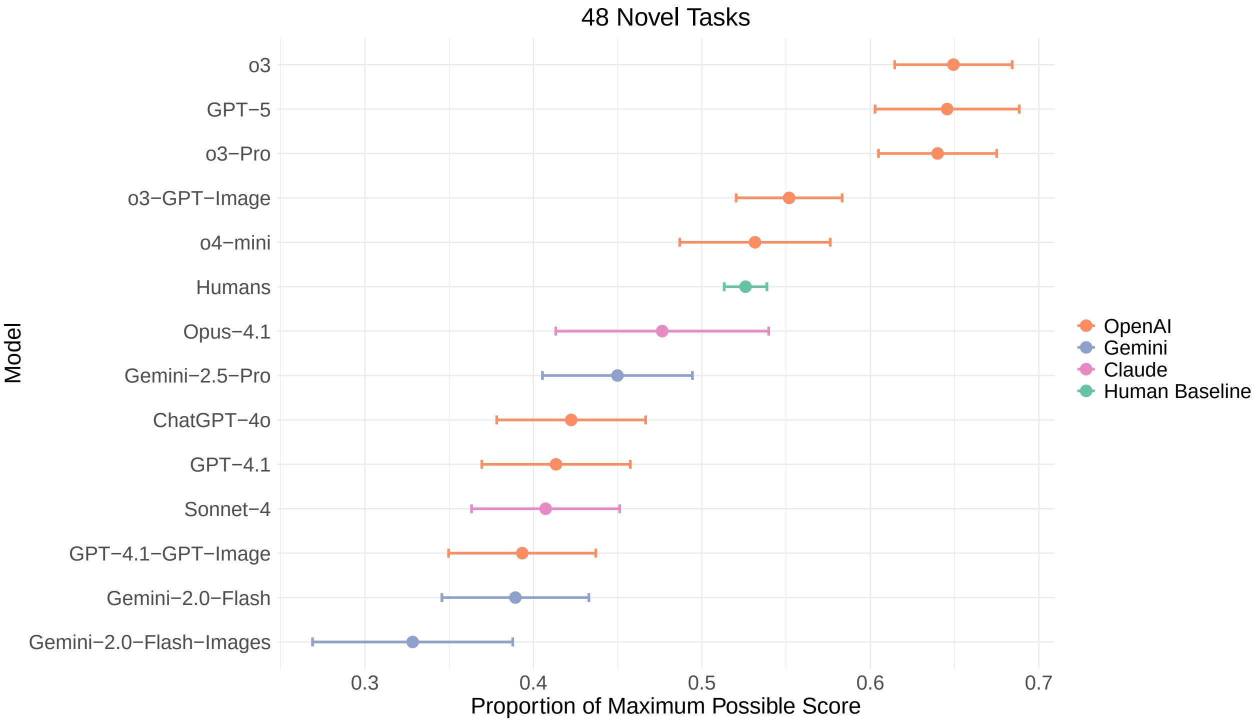
```
##
##
##                comparison                                        diff   chi_squared   p_value  signifi
## ------------  ---------------------------------------------  -------  ------------  --------  -------
## X-squared     Humans vs o3                                    -0.123    38.8606815    0.0000  TRUE
## X-squared1    Humans vs o3-GPT-Image                          -0.026     2.1151564    0.1458  FALSE
## X-squared2    Humans vs o3-Pro                                -0.114    33.1116612    0.0000  TRUE
## X-squared3    Humans vs GPT-4.1                                0.112    22.0592775    0.0000  TRUE
## X-squared4    Humans vs GPT-4.1-GPT-Image                      0.133    30.7158709    0.0000  TRUE
## X-squared5    Humans vs ChatGPT-4o                             0.103    18.6444291    0.0000  TRUE
## X-squared6    Humans vs o4-mini                               -0.006     0.0354255    0.8507  FALSE
## X-squared7    Humans vs Gemini-2.5-Pro                         0.076     9.9860525    0.0016  TRUE
## X-squared8    Humans vs Gemini-2.0-Flash                       0.137    32.6384035    0.0000  TRUE
## X-squared9    Humans vs Gemini-2.0-Flash-Images                0.198    35.2801051    0.0000  TRUE
## X-squared10   Humans vs Sonnet-4                               0.119    24.5745389    0.0000  TRUE
## X-squared11   Humans vs Opus-4.1                               0.049     2.0682572    0.1504  FALSE
## X-squared12   Humans vs GPT-5                                 -0.120    25.0838266    0.0000  TRUE
## X-squared13   o3 vs o3-GPT-Image                               0.098    15.8181217    0.0001  TRUE
## X-squared14   o3 vs o3-Pro                                     0.009     0.1011775    0.7504  FALSE
## X-squared15   o3 vs GPT-4.1                                    0.236    63.9096051    0.0000  TRUE
## X-squared16   o3 vs GPT-4.1-GPT-Image                          0.256    75.0810006    0.0000  TRUE
## X-squared17   o3 vs ChatGPT-4o                                 0.227    59.1955498    0.0000  TRUE
## X-squared18   o3 vs o4-mini                                    0.118    16.1910358    0.0001  TRUE
## X-squared19   o3 vs Gemini-2.5-Pro                             0.199    45.8969507    0.0000  TRUE
## X-squared20   o3 vs Gemini-2.0-Flash                           0.260    77.4476316    0.0000  TRUE
## X-squared21   o3 vs Gemini-2.0-Flash-Images                    0.321    74.2960065    0.0000  TRUE
## X-squared22   o3 vs Sonnet-4                                   0.242    67.2562927    0.0000  TRUE
## X-squared23   o3 vs Opus-4.1                                   0.173    21.8007438    0.0000  TRUE
## X-squared24  o3 vs GPT-5                                    0.004    0.0049901   0.9437  FALSE
## X-squared25  o3-GPT-Image vs o3-Pro                        -0.088   12.8378194   0.0003  TRUE
## X-squared26  o3-GPT-Image vs GPT-4.1                        0.138   23.9431768   0.0000  TRUE
## X-squared27  o3-GPT-Image vs GPT-4.1-GPT-Image              0.158   31.4772295   0.0000  TRUE
## X-squared28  o3-GPT-Image vs ChatGPT-4o                     0.129   20.9038155   0.0000  TRUE
## X-squared29  o3-GPT-Image vs o4-mini                        0.020    0.4513398   0.5017  FALSE
## X-squared30  o3-GPT-Image vs Gemini-2.5-Pro                 0.102   12.8968083   0.0003  TRUE
## X-squared31  o3-GPT-Image vs Gemini-2.0-Flash               0.162   33.1257777   0.0000  TRUE
## X-squared32  o3-GPT-Image vs Gemini-2.0-Flash-Images        0.223   37.4503448   0.0000  TRUE
## X-squared33  o3-GPT-Image vs Sonnet-4                       0.145   26.1542007   0.0000  TRUE
## X-squared34  o3-GPT-Image vs Opus-4.1                       0.075    4.0807836   0.0434  TRUE
## X-squared35  o3-GPT-Image vs GPT-5                         -0.094   11.1983901   0.0008  TRUE
## X-squared36  o3-Pro vs GPT-4.1                              0.226   58.7335492   0.0000  TRUE
## X-squared37  o3-Pro vs GPT-4.1-GPT-Image                    0.246   69.4830297   0.0000  TRUE
## X-squared38  o3-Pro vs ChatGPT-4o                           0.217   54.2103917   0.0000  TRUE
## X-squared39  o3-Pro vs o4-mini                              0.108   13.6072593   0.0002  TRUE
## X-squared40  o3-Pro vs Gemini-2.5-Pro                       0.190   41.5021036   0.0000  TRUE
## X-squared41  o3-Pro vs Gemini-2.0-Flash                     0.251   71.7651322   0.0000  TRUE
## X-squared42  o3-Pro vs Gemini-2.0-Flash-Images              0.312   69.7017215   0.0000  TRUE
## X-squared43  o3-Pro vs Sonnet-4                             0.233   61.9496000   0.0000  TRUE
## X-squared44  o3-Pro vs Opus-4.1                             0.163   19.3366750   0.0000  TRUE
## X-squared45  o3-Pro vs GPT-5                               -0.006    0.0195799   0.8887  FALSE
## X-squared46  GPT-4.1 vs GPT-4.1-GPT-Image                   0.020    0.3221521   0.5703  FALSE
## X-squared47  GPT-4.1 vs ChatGPT-4o                         -0.009    0.0470018   0.8284  FALSE
## X-squared48  GPT-4.1 vs o4-mini                            -0.118   12.9513714   0.0003  TRUE
## X-squared49  GPT-4.1 vs Gemini-2.5-Pro                     -0.036    1.1551296   0.2825  FALSE
## X-squared50  GPT-4.1 vs Gemini-2.0-Flash                    0.024    0.4845990   0.4863  FALSE
## X-squared51  GPT-4.1 vs Gemini-2.0-Flash-Images             0.085    4.5388375   0.0331  TRUE
## X-squared52  GPT-4.1 vs Sonnet-4                            0.006    0.0166510   0.8973  FALSE
## X-squared53  GPT-4.1 vs Opus-4.1                           -0.063    2.3361536   0.1264  FALSE
## X-squared54  GPT-4.1 vs GPT-5                              -0.232   50.9779767   0.0000  TRUE
## X-squared55  GPT-4.1-GPT-Image vs ChatGPT-4o               -0.029    0.7222319   0.3954  FALSE
## X-squared56  GPT-4.1-GPT-Image vs o4-mini                  -0.138   17.8652217   0.0000  TRUE
## X-squared57  GPT-4.1-GPT-Image vs Gemini-2.5-Pro           -0.057    2.9146207   0.0878  FALSE
## X-squared58  GPT-4.1-GPT-Image vs Gemini-2.0-Flash          0.004    0.0039073   0.9502  FALSE
## X-squared59  GPT-4.1-GPT-Image vs Gemini-2.0-Flash-Images   0.065    2.6252260   0.1052  FALSE
## X-squared60  GPT-4.1-GPT-Image vs Sonnet-4                 -0.014    0.1390405   0.7092  FALSE
## X-squared61  GPT-4.1-GPT-Image vs Opus-4.1                 -0.083    4.1956974   0.0405  TRUE
## X-squared62  GPT-4.1-GPT-Image vs GPT-5                    -0.252   60.1365189   0.0000  TRUE
## X-squared63  ChatGPT-4o vs o4-mini                         -0.109   11.0121623   0.0009  TRUE
## X-squared64  ChatGPT-4o vs Gemini-2.5-Pro                  -0.027    0.6284762   0.4279  FALSE
## X-squared65  ChatGPT-4o vs Gemini-2.0-Flash                 0.033    0.9571819   0.3279  FALSE
## X-squared66  ChatGPT-4o vs Gemini-2.0-Flash-Images          0.094    5.5571569   0.0184  TRUE
## X-squared67  ChatGPT-4o vs Sonnet-4                         0.015    0.1692220   0.6808  FALSE
## X-squared68  ChatGPT-4o vs Opus-4.1                        -0.054    1.6826566   0.1946  FALSE
## X-squared69  ChatGPT-4o vs GPT-5                           -0.223   47.1271745   0.0000  TRUE
## X-squared70  o4-mini vs Gemini-2.5-Pro                      0.082    6.0739338   0.0137  TRUE
## X-squared71  o4-mini vs Gemini-2.0-Flash                    0.142   18.9574569   0.0000  TRUE
## X-squared72  o4-mini vs Gemini-2.0-Flash-Images             0.203   25.7394291   0.0000  TRUE
## X-squared73  o4-mini vs Sonnet-4                            0.124   14.3789349   0.0001  TRUE
## X-squared74  o4-mini vs Opus-4.1                            0.055    1.7259167   0.1889  FALSE
## X-squared75  o4-mini vs GPT-5                              -0.114   12.4274222   0.0004  TRUE
## X-squared76  Gemini-2.5-Pro vs Gemini-2.0-Flash             0.061    3.3693265   0.0664  FALSE
## X-squared77  Gemini-2.5-Pro vs Gemini-2.0-Flash-Images      0.122    9.2773966   0.0023  TRUE
## X-squared78   Gemini-2.5-Pro vs Sonnet-4                       0.043     1.6108643    0.2044  FALSE
## X-squared79   Gemini-2.5-Pro vs Opus-4.1                      -0.027     0.3543462    0.5517  FALSE
## X-squared80   Gemini-2.5-Pro vs GPT-5                         -0.196    36.3093587    0.0000  TRUE
## X-squared81   Gemini-2.0-Flash vs Gemini-2.0-Flash-Images      0.061     2.2996419    0.1294  FALSE
## X-squared82   Gemini-2.0-Flash vs Sonnet-4                    -0.018     0.2514460    0.6161  FALSE
## X-squared83   Gemini-2.0-Flash vs Opus-4.1                    -0.087     4.6400681    0.0312  TRUE
## X-squared84   Gemini-2.0-Flash vs GPT-5                       -0.256    62.0826267    0.0000  TRUE
## X-squared85   Gemini-2.0-Flash-Images vs Sonnet-4             -0.079     3.8959604    0.0484  TRUE
## X-squared86   Gemini-2.0-Flash-Images vs Opus-4.1             -0.148    10.3385839    0.0013  TRUE
## X-squared87   Gemini-2.0-Flash-Images vs GPT-5                -0.317    63.5367890    0.0000  TRUE
## X-squared88   Sonnet-4 vs Opus-4.1                            -0.069     2.8496179    0.0914  FALSE
## X-squared89   Sonnet-4 vs GPT-5                               -0.238    53.7168275    0.0000  TRUE
## X-squared90   Opus-4.1 vs GPT-5                               -0.169    18.2100464    0.0000  TRUE
```

Visualization of All Comparisons

```r
# 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, aes(color = color)) +
  geom_errorbar(aes(ymin = proportion - 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    ymax = proportion + 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    color = color),
                width = 0.2, size = 1) +
  coord_flip() +
  theme_minimal() +
  labs(subtitle = "Finke et al. Sets",
       x = "Model",
       y = "Proportion of Maximum Possible Score") +
  theme(plot.subtitle = element_text(hjust = 0.5, size = 18),
        axis.text = element_text(size = 14),
        axis.title = element_text(size = 16),
        legend.text = element_text(size = 14)) +
  scale_color_manual(
    values = c("#fc8d62", "#8da0cb", "#e78ac3", "#66c2a5"),
    name = element_blank(),
    breaks = c("#fc8d62", "#8da0cb", "#e78ac3", "#66c2a5"),
    labels = c("OpenAI", "Gemini", "Claude", "Human Baseline")
  )

# 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, aes(color = color)) +
  geom_errorbar(aes(ymin = proportion - 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    ymax = proportion + 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    color = color),
                width = 0.2, size = 1) +
  coord_flip() +
  theme_minimal() +
  labs(subtitle = "Novel 48-Item Sets",
       x = "Model",
       y = "Proportion of Maximum Possible Score") +
  theme(plot.subtitle = element_text(hjust = 0.5, size = 18),
        axis.text = element_text(size = 14),
        axis.title = element_text(size = 16),
        legend.text = element_text(size = 14)) +
  scale_color_manual(
    values = c("#fc8d62", "#8da0cb", "#e78ac3", "#66c2a5"),
    name = element_blank(),
    breaks = c("#fc8d62", "#8da0cb", "#e78ac3", "#66c2a5"),
    labels = c("OpenAI", "Gemini", "Claude", "Human Baseline")
  )

# Combine plots
combined_plot <- ((finke_plot + novel_48_plot) +
  plot_layout(ncol = 2, guides = "collect") +
  plot_annotation(title = "Finke et al. Tasks vs. 48 Novel Tasks - Proportions with 95% CI")) &
  theme(plot.title = element_text(hjust = 0.5, size = 20, face = "bold"), legend.position = "bottom")
print(combined_plot)
```



```
novel_plot <- ggplot(novel_data, aes(x = reorder(model, proportion), y = proportion)) +
  geom_point(size = 4, aes(color = color)) +
  geom_errorbar(aes(ymin = proportion - 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    ymax = proportion + 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    color = color),
                width = 0.2, size = 1) +
  coord_flip() +
  theme_minimal() +
  labs(subtitle = "48 Novel Tasks",
       x = "Model",
       y = "Proportion of Maximum Possible Score") +
  theme(plot.subtitle = element_text(hjust = 0.5, size = 20),
        axis.text = element_text(size = 16),
        axis.title = element_text(size = 18),
        legend.text = element_text(size = 16)) +
  scale_color_manual(
    values = c("#fc8d62", "#8da0cb", "#e78ac3", "#66c2a5"),
    name = element_blank(),
    breaks = c("#fc8d62", "#8da0cb", "#e78ac3", "#66c2a5"),
    labels = c("OpenAI", "Gemini", "Claude", "Human Baseline")
  )
print(novel_plot)
```

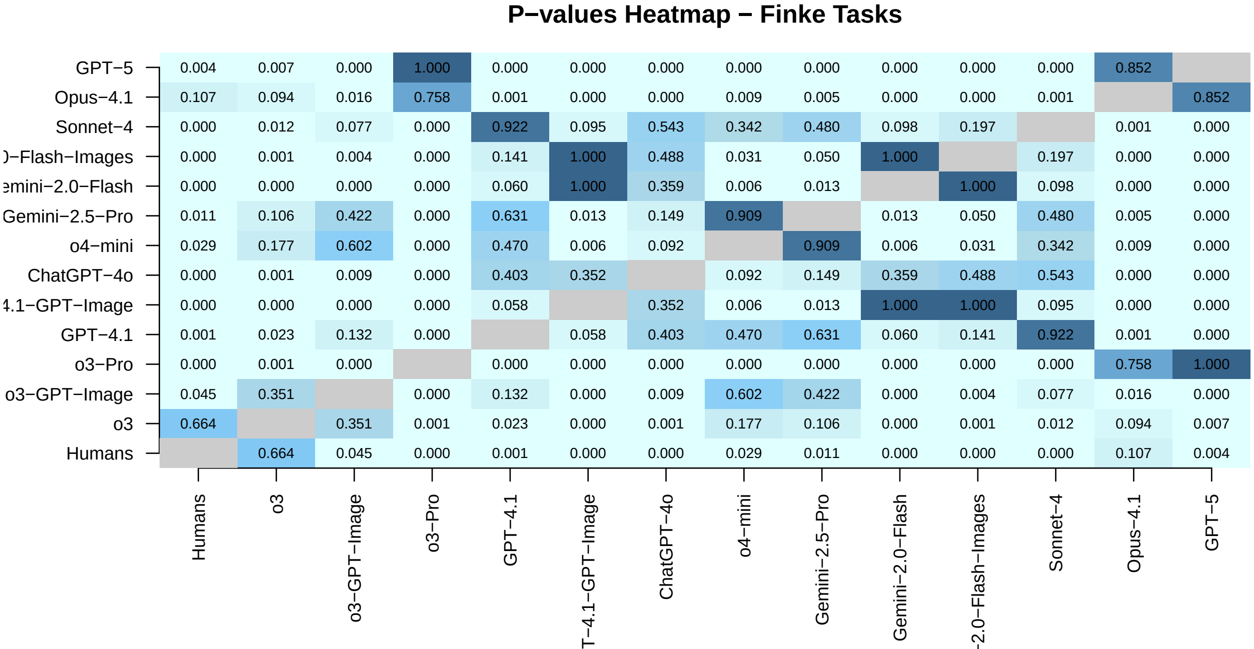


Heatmap of P-values

```
# 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
```

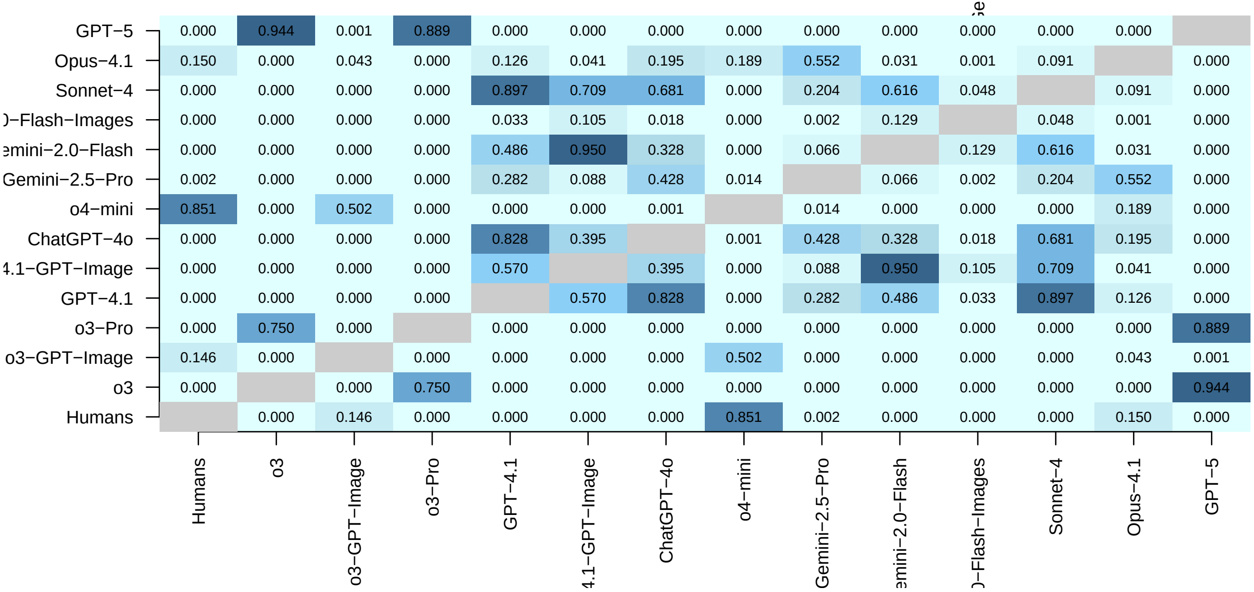
```r
for (i in 1:nrow(novel_48_results)) {
  row_idx <- which(novel_models == novel_48_results$model1[i])
  col_idx <- which(novel_models == novel_48_results$model2[i])
  novel_pval_matrix[row_idx, col_idx] <- novel_48_results$p_value[i]
  novel_pval_matrix[col_idx, row_idx] <- novel_48_results$p_value[i]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

# 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−Pro	Gemini−2.0−Flash	emini−2.0−Flash−Images	Sonnet−4	Opus−4.1	GPT−5
GPT−5	0.004	0.007	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.852	
Opus−4.1	0.107	0.094	0.016	0.758	0.001	0.000	0.000	0.009	0.005	0.000	0.000	0.001		0.852
Sonnet−4	0.000	0.012	0.077	0.000	0.922	0.095	0.543	0.342	0.480	0.098	0.197		0.001	0.000
)−Flash−Images	0.000	0.001	0.004	0.000	0.141	1.000	0.488	0.031	0.050	1.000		0.197	0.000	0.000
emini−2.0−Flash	0.000	0.000	0.000	0.000	0.060	1.000	0.359	0.006	0.013		1.000	0.098	0.000	0.000
Gemini−2.5−Pro	0.011	0.106	0.422	0.000	0.631	0.013	0.149	0.909		0.013	0.050	0.480	0.005	0.000
o4−mini	0.029	0.177	0.602	0.000	0.470	0.006	0.092		0.909	0.006	0.031	0.342	0.009	0.000
ChatGPT−4o	0.000	0.001	0.009	0.000	0.403	0.352		0.092	0.149	0.359	0.488	0.543	0.000	0.000
4.1−GPT−Image	0.000	0.000	0.000	0.000	0.058		0.352	0.006	0.013	1.000	1.000	0.095	0.000	0.000
GPT−4.1	0.001	0.023	0.132	0.000		0.058	0.403	0.470	0.631	0.060	0.141	0.922	0.001	0.000
o3−Pro	0.000	0.001	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.758	1.000
o3−GPT−Image	0.045	0.351		0.000	0.132	0.000	0.009	0.602	0.422	0.000	0.004	0.077	0.016	0.000
o3	0.664		0.351	0.001	0.023	0.000	0.001	0.177	0.106	0.000	0.001	0.012	0.094	0.007
Humans		0.664	0.045	0.000	0.001	0.000	0.000	0.029	0.011	0.000	0.000	0.000	0.107	0.004

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−Pro	emini−2.0−Flash	)−Flash−Images	Sonnet−4	Opus−4.1	GPT−5
GPT−5	0.000	0.944	0.001	0.889	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
Opus−4.1	0.150	0.000	0.043	0.000	0.126	0.041	0.195	0.189	0.552	0.031	0.001	0.091		0.000
Sonnet−4	0.000	0.000	0.000	0.000	0.897	0.709	0.681	0.000	0.204	0.616	0.048		0.091	0.000
)−Flash−Images	0.000	0.000	0.000	0.000	0.033	0.105	0.018	0.000	0.002	0.129		0.048	0.001	0.000
emini−2.0−Flash	0.000	0.000	0.000	0.000	0.486	0.950	0.328	0.000	0.066		0.129	0.616	0.031	0.000
Gemini−2.5−Pro	0.002	0.000	0.000	0.000	0.282	0.088	0.428	0.014		0.066	0.002	0.204	0.552	0.000
o4−mini	0.851	0.000	0.502	0.000	0.000	0.000	0.001		0.014	0.000	0.000	0.000	0.189	0.000
ChatGPT−4o	0.000	0.000	0.000	0.000	0.828	0.395		0.001	0.428	0.328	0.018	0.681	0.195	0.000
4.1−GPT−Image	0.000	0.000	0.000	0.000	0.570		0.395	0.000	0.088	0.950	0.105	0.709	0.041	0.000
GPT−4.1	0.000	0.000	0.000	0.000		0.570	0.828	0.000	0.282	0.486	0.033	0.897	0.126	0.000
o3−Pro	0.000	0.750	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.889
o3−GPT−Image	0.146	0.000		0.000	0.000	0.000	0.000	0.502	0.000	0.000	0.000	0.000	0.043	0.001
o3	0.000		0.000	0.750	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.944
Humans		0.000	0.146	0.000	0.000	0.000	0.000	0.851	0.002	0.000	0.000	0.000	0.150	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")
```

```
## ==================================================
cat("Finke Tasks:\n")
```

```
## Finke Tasks:
```

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

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

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

```
##   Significant differences: 56
```

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

```
##   Percentage significant: 61.5 %
```

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

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

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

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

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

```
##   Significant differences: 62
```

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

```
##   Percentage significant: 68.1 %
```

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

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

```
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-squared1    Humans vs o3-GPT-Image                       0.0699    0.0452
## X-squared2    Humans vs o3-Pro                            -0.1415    0.0002
## X-squared3    Humans vs GPT-4.1                            0.1602    0.0007
## X-squared4    Humans vs GPT-4.1-GPT-Image                  0.2886    0.0000
## X-squared5    Humans vs ChatGPT-4o                         0.2221    0.0000
## X-squared6    Humans vs o4-mini                            0.1052    0.0285
## X-squared7    Humans vs Gemini-2.5-Pro                     0.1209    0.0114
## X-squared8    Humans vs Gemini-2.0-Flash                   0.2877    0.0000
## X-squared9    Humans vs Gemini-2.0-Flash-Images            0.2879    0.0000
## X-squared10   Humans vs Sonnet-4                           0.1748    0.0002
## X-squared12   Humans vs GPT-5                             -0.1360    0.0038
## X-squared14   o3 vs o3-Pro                                -0.1612    0.0014
## X-squared15   o3 vs GPT-4.1                                0.1405    0.0226
## X-squared16   o3 vs GPT-4.1-GPT-Image                       0.2689    0.0000
## X-squared17   o3 vs ChatGPT-4o                              0.2024    0.0009
## X-squared20   o3 vs Gemini-2.0-Flash                        0.2681    0.0000
## X-squared21   o3 vs Gemini-2.0-Flash-Images                 0.2683    0.0005
## X-squared22   o3 vs Sonnet-4                                0.1551    0.0115
## X-squared24   o3 vs GPT-5                                  -0.1557    0.0071
## X-squared25   o3-GPT-Image vs o3-Pro                       -0.2114    0.0000
## X-squared27   o3-GPT-Image vs GPT-4.1-GPT-Image             0.2187    0.0001
## X-squared28   o3-GPT-Image vs ChatGPT-4o                    0.1522    0.0090
## X-squared31   o3-GPT-Image vs Gemini-2.0-Flash              0.2178    0.0002
## X-squared32   o3-GPT-Image vs Gemini-2.0-Flash-Images       0.2180    0.0040
## X-squared34   o3-GPT-Image vs Opus-4.1                     -0.1808    0.0161
## X-squared35   o3-GPT-Image vs GPT-5                        -0.2059    0.0002
## X-squared36   o3-Pro vs GPT-4.1                             0.3017    0.0000
## X-squared37   o3-Pro vs GPT-4.1-GPT-Image                   0.4300    0.0000
## X-squared38   o3-Pro vs ChatGPT-4o                          0.3635    0.0000
## X-squared39   o3-Pro vs o4-mini                             0.2466    0.0000
## X-squared40   o3-Pro vs Gemini-2.5-Pro                      0.2623    0.0000
## X-squared41   o3-Pro vs Gemini-2.0-Flash                    0.4292    0.0000
## X-squared42   o3-Pro vs Gemini-2.0-Flash-Images             0.4294    0.0000
## X-squared43   o3-Pro vs Sonnet-4                            0.3163    0.0000
## X-squared53   GPT-4.1 vs Opus-4.1                          -0.2711    0.0010
## X-squared54   GPT-4.1 vs GPT-5                             -0.2962    0.0000
## X-squared56   GPT-4.1-GPT-Image vs o4-mini                 -0.1834    0.0062
## X-squared57   GPT-4.1-GPT-Image vs Gemini-2.5-Pro          -0.1677    0.0125
## X-squared61   GPT-4.1-GPT-Image vs Opus-4.1                -0.3994    0.0000
## X-squared62   GPT-4.1-GPT-Image vs GPT-5                   -0.4246    0.0000
## X-squared68   ChatGPT-4o vs Opus-4.1                       -0.3329    0.0000
## X-squared69   ChatGPT-4o vs GPT-5                          -0.3581    0.0000
## X-squared71   o4-mini vs Gemini-2.0-Flash                   0.1826    0.0065
## X-squared72   o4-mini vs Gemini-2.0-Flash-Images            0.1828    0.0308
## X-squared74   o4-mini vs Opus-4.1                          -0.2160    0.0087
## X-squared75   o4-mini vs GPT-5                             -0.2412    0.0002
## X-squared76   Gemini-2.5-Pro vs Gemini-2.0-Flash            0.1669    0.0130
## X-squared77   Gemini-2.5-Pro vs Gemini-2.0-Flash-Images     0.1671    0.0496
## X-squared79   Gemini-2.5-Pro vs Opus-4.1                   -0.2317    0.0049
## X-squared80   Gemini-2.5-Pro vs GPT-5                      -0.2569    0.0001
## X-squared83   Gemini-2.0-Flash vs Opus-4.1                 -0.3986    0.0000
## X-squared84   Gemini-2.0-Flash vs GPT-5                    -0.4237    0.0000
## X-squared86   Gemini-2.0-Flash-Images vs Opus-4.1          -0.3988    0.0000
## X-squared87   Gemini-2.0-Flash-Images vs GPT-5             -0.4239    0.0000
## X-squared88   Sonnet-4 vs Opus-4.1                         -0.2857    0.0005
## X-squared89   Sonnet-4 vs GPT-5                            -0.3108    0.0000
```

```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                                 -0.1234    0.0000
## X-squared2    Humans vs o3-Pro                             -0.1140    0.0000
## X-squared3    Humans vs GPT-4.1                             0.1125    0.0000
## X-squared4    Humans vs GPT-4.1-GPT-Image                   0.1325    0.0000
## X-squared5    Humans vs ChatGPT-4o                          0.1035    0.0000
## X-squared7    Humans vs Gemini-2.5-Pro                      0.0760    0.0016
## X-squared8    Humans vs Gemini-2.0-Flash                    0.1366    0.0000
## X-squared9    Humans vs Gemini-2.0-Flash-Images             0.1976    0.0000
## X-squared10   Humans vs Sonnet-4                            0.1187    0.0000
## X-squared12   Humans vs GPT-5                              -0.1196    0.0000
## X-squared13   o3 vs o3-GPT-Image                            0.0975    0.0001
## X-squared15   o3 vs GPT-4.1                                 0.2358    0.0000
## X-squared16   o3 vs GPT-4.1-GPT-Image                       0.2559    0.0000
## X-squared17   o3 vs ChatGPT-4o                              0.2269    0.0000
## X-squared18   o3 vs o4-mini                                 0.1178    0.0001
## X-squared19   o3 vs Gemini-2.5-Pro                          0.1994    0.0000
## X-squared20   o3 vs Gemini-2.0-Flash                        0.2600    0.0000
## X-squared21   o3 vs Gemini-2.0-Flash-Images                 0.3209    0.0000
## X-squared22   o3 vs Sonnet-4                                0.2420    0.0000
## X-squared23   o3 vs Opus-4.1                                0.1728    0.0000
## X-squared25   o3-GPT-Image vs o3-Pro                       -0.0881    0.0003
## X-squared26   o3-GPT-Image vs GPT-4.1                       0.1383    0.0000
## X-squared27   o3-GPT-Image vs GPT-4.1-GPT-Image             0.1584    0.0000
## X-squared28   o3-GPT-Image vs ChatGPT-4o                    0.1293    0.0000
## X-squared30   o3-GPT-Image vs Gemini-2.5-Pro                0.1019    0.0003
## X-squared31   o3-GPT-Image vs Gemini-2.0-Flash              0.1624    0.0000
## X-squared32   o3-GPT-Image vs Gemini-2.0-Flash-Images       0.2234    0.0000
## X-squared33   o3-GPT-Image vs Sonnet-4                      0.1445    0.0000
## X-squared34   o3-GPT-Image vs Opus-4.1                      0.0753    0.0434
## X-squared35   o3-GPT-Image vs GPT-5                        -0.0938    0.0008
## X-squared36   o3-Pro vs GPT-4.1                             0.2264    0.0000
## X-squared37   o3-Pro vs GPT-4.1-GPT-Image                   0.2465    0.0000
## X-squared38   o3-Pro vs ChatGPT-4o                          0.2174    0.0000
## X-squared39   o3-Pro vs o4-mini                             0.1084    0.0002
## X-squared40   o3-Pro vs Gemini-2.5-Pro                      0.1900    0.0000
## X-squared41   o3-Pro vs Gemini-2.0-Flash                    0.2505    0.0000
## X-squared42   o3-Pro vs Gemini-2.0-Flash-Images             0.3115    0.0000
## X-squared43   o3-Pro vs Sonnet-4                            0.2326    0.0000
## X-squared44   o3-Pro vs Opus-4.1                            0.1634    0.0000
## X-squared48   GPT-4.1 vs o4-mini                           -0.1181    0.0003
## X-squared51   GPT-4.1 vs Gemini-2.0-Flash-Images            0.0851    0.0331
## X-squared54   GPT-4.1 vs GPT-5                             -0.2321    0.0000
## X-squared56   GPT-4.1-GPT-Image vs o4-mini                 -0.1381    0.0000
## X-squared61   GPT-4.1-GPT-Image vs Opus-4.1                -0.0831    0.0405
## X-squared62   GPT-4.1-GPT-Image vs GPT-5                   -0.2522    0.0000
## X-squared63   ChatGPT-4o vs o4-mini                        -0.1091    0.0009
## X-squared66   ChatGPT-4o vs Gemini-2.0-Flash-Images         0.0941    0.0184
## X-squared69   ChatGPT-4o vs GPT-5                          -0.2231    0.0000
## X-squared70   o4-mini vs Gemini-2.5-Pro                     0.0816    0.0137
## X-squared71   o4-mini vs Gemini-2.0-Flash                   0.1422    0.0000
## X-squared72   o4-mini vs Gemini-2.0-Flash-Images           0.2031    0.0000
## X-squared73   o4-mini vs Sonnet-4                          0.1242    0.0001
## X-squared75   o4-mini vs GPT-5                            -0.1141    0.0004
## X-squared77   Gemini-2.5-Pro vs Gemini-2.0-Flash-Images    0.1215    0.0023
## X-squared80   Gemini-2.5-Pro vs GPT-5                     -0.1957    0.0000
## X-squared83   Gemini-2.0-Flash vs Opus-4.1                -0.0871    0.0312
## X-squared84   Gemini-2.0-Flash vs GPT-5                   -0.2562    0.0000
## X-squared85   Gemini-2.0-Flash-Images vs Sonnet-4         -0.0789    0.0484
## X-squared86   Gemini-2.0-Flash-Images vs Opus-4.1         -0.1481    0.0013
## X-squared87   Gemini-2.0-Flash-Images vs GPT-5            -0.3172    0.0000
## X-squared89   Sonnet-4 vs GPT-5                           -0.2383    0.0000
## X-squared90   Opus-4.1 vs GPT-5                           -0.1691    0.0000
```

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:  0.547  vs  0.642
## Difference:  -0.094
## Chi-squared:  28.631
## Degrees of freedom:  1
## P-value:  0.00000008757
## 95% CI: [ -0.128 ,  -0.06 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o3-GPT-Image
## ----------------------------------------
## Proportions:  0.547  vs  0.553
## Difference:  -0.006
## Chi-squared:  0.143
## Degrees of freedom:  1
## P-value:  0.7057
## 95% CI: [ -0.037 ,  0.024 ]
## Significant:  NO
##
##  Humans vs o3-Pro
## ----------------------------------------
## Proportions:  0.547  vs  0.666
## Difference:  -0.119
## Chi-squared:  45.76
## Degrees of freedom:  1
## P-value:  0.00000000001336
## 95% CI: [ -0.153 ,  -0.086 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-4.1
## ----------------------------------------
## Proportions:  0.547  vs  0.425
## Difference:  0.122
## Chi-squared:  32.987
## Degrees of freedom:  1
## P-value:  0.000000009278
## 95% CI: [ 0.08 ,  0.164 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.547  vs  0.383
## Difference:  0.164
## Chi-squared:  59.482
## Degrees of freedom:  1
## P-value:  0.00000000000001234
## 95% CI: [ 0.123 ,  0.206 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.547  vs  0.42
## Difference:  0.128
## Chi-squared:  35.86
## Degrees of freedom:  1
## P-value:  0.00000000212
## 95% CI: [ 0.086 ,  0.17 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o4-mini
## ----------------------------------------
## Proportions:  0.547  vs  0.53
## Difference:  0.017
## Chi-squared:  0.577
## Degrees of freedom:  1
## P-value:  0.4477
## 95% CI: [ -0.025 ,  0.059 ]
## Significant:  NO
##
##  Humans vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.547  vs  0.462
## Difference:  0.085
## Chi-squared:  15.96
## Degrees of freedom:  1
## P-value:  0.00006468
## 95% CI: [ 0.043 ,  0.128 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.547  vs  0.38
## Difference:  0.167
## Chi-squared:  61.741
## Degrees of freedom:  1
## P-value:  0.000000000000003918
## 95% CI: [ 0.126 ,  0.209 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.547  vs  0.331
## Difference:  0.216
## Chi-squared:  53.296
## Degrees of freedom:  1
## P-value:  0.0000000000002869
## 95% CI: [ 0.16 ,  0.272 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Sonnet-4
## ----------------------------------------
## Proportions:  0.547  vs  0.417
## Difference:  0.13
## Chi-squared:  37.392
## Degrees of freedom:  1
## P-value:  0.0000000009662
## 95% CI: [ 0.088 ,  0.172 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Opus-4.1
## ----------------------------------------
## Proportions:  0.547  vs  0.529
## Difference:  0.018
## Chi-squared:  0.299
## Degrees of freedom:  1
## P-value:  0.5845
## 95% CI: [ -0.042 ,  0.077 ]
## Significant:  NO
##
##  Humans vs GPT-5
## ----------------------------------------
## Proportions:  0.547  vs  0.67
## Difference:  -0.123
## Chi-squared:  33.302
## Degrees of freedom:  1
## P-value:  0.00000000789
## 95% CI: [ -0.163 ,  -0.082 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o3-GPT-Image
## ----------------------------------------
## Proportions:  0.642  vs  0.553
## Difference:  0.088
## Chi-squared:  16.139
## Degrees of freedom:  1
## P-value:  0.00005886
## 95% CI: [ 0.045 ,  0.131 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o3-Pro
## ----------------------------------------
## Proportions:  0.642  vs  0.666
## Difference:  -0.025
## Chi-squared:  1.106
## Degrees of freedom:  1
## P-value:  0.2928
## 95% CI: [ -0.07 ,  0.02 ]
## Significant:  NO
##
##  o3 vs GPT-4.1
## ----------------------------------------
## Proportions:  0.642  vs  0.425
## Difference:  0.217
## Chi-squared:  67.617
## Degrees of freedom:  1
## P-value:  0.0000000000000001986
## 95% CI: [ 0.165 ,  0.269 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.642  vs  0.383
## Difference:  0.258
## Chi-squared:  95.75
## Degrees of freedom:  1
## P-value:  0.0000000000000000000001304
## 95% CI: [ 0.207 ,  0.31 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.642  vs  0.42
## Difference:  0.222
## Chi-squared:  70.849
## Degrees of freedom:  1
## P-value:  0.00000000000000003855
## 95% CI: [ 0.17 ,  0.274 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o4-mini
## ----------------------------------------
## Proportions:  0.642  vs  0.53
## Difference:  0.111
## Chi-squared:  18.085
## Degrees of freedom:  1
## P-value:  0.00002112
## 95% CI: [ 0.059 ,  0.163 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.642  vs  0.462
## Difference:  0.18
## Chi-squared:  46.719
## Degrees of freedom:  1
## P-value:  0.000000000008192
## 95% CI: [ 0.128 ,  0.232 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.642  vs  0.38
## Difference:  0.262
## Chi-squared:  98.017
## Degrees of freedom:  1
## P-value:  0.0000000000000000000004147
## 95% CI: [ 0.21 ,  0.313 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.642  vs  0.331
## Difference:  0.31
## Chi-squared:  86.894
## Degrees of freedom:  1
## P-value:  0.00000000000000000001145
## 95% CI: [ 0.246 ,  0.374 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Sonnet-4
## ----------------------------------------
## Proportions:  0.642  vs  0.417
## Difference:  0.225
## Chi-squared:  72.549
## Degrees of freedom:  1
## P-value:  0.0000000000000001629
## 95% CI: [ 0.173 ,  0.276 ]
## Significant:  YES (p < 0.05)
```

```
##
##  o3 vs Opus-4.1
## ----------------------------------------
## Proportions:  0.642  vs  0.529
## Difference:  0.112
## Chi-squared:  11.466
## Degrees of freedom:  1
## P-value:  0.0007087
## 95% CI: [ 0.045 ,  0.179 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs GPT-5
## ----------------------------------------
## Proportions:  0.642  vs  0.67
## Difference:  -0.028
## Chi-squared:  1.138
## Degrees of freedom:  1
## P-value:  0.286
## 95% CI: [ -0.079 ,  0.022 ]
## Significant:  NO
##
##  o3-GPT-Image vs o3-Pro
## ----------------------------------------
## Proportions:  0.553  vs  0.666
## Difference:  -0.113
## Chi-squared:  26.82
## Degrees of freedom:  1
## P-value:  0.0000002234
## 95% CI: [ -0.155 ,  -0.07 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs GPT-4.1
## ----------------------------------------
## Proportions:  0.553  vs  0.425
## Difference:  0.129
## Chi-squared:  26.007
## Degrees of freedom:  1
## P-value:  0.0000003401
## 95% CI: [ 0.079 ,  0.178 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.553  vs  0.383
## Difference:  0.17
## Chi-squared:  45.802
## Degrees of freedom:  1
## P-value:  0.00000000001308
## 95% CI: [ 0.121 ,  0.22 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.553  vs  0.42
## Difference:  0.134
## Chi-squared:  28.164
## Degrees of freedom:  1
## P-value:  0.0000001115
## 95% CI: [ 0.084 ,  0.184 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs o4-mini
## ----------------------------------------
## Proportions:  0.553  vs  0.53
## Difference:  0.023
## Chi-squared:  0.782
## Degrees of freedom:  1
## P-value:  0.3765
## 95% CI: [ -0.027 ,  0.073 ]
## Significant:  NO
##
##  o3-GPT-Image vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.553  vs  0.462
## Difference:  0.092
## Chi-squared:  13.116
## Degrees of freedom:  1
## P-value:  0.0002928
## 95% CI: [ 0.042 ,  0.142 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.553  vs  0.38
## Difference:  0.174
## Chi-squared:  47.484
## Degrees of freedom:  1
## P-value:  0.000000000005546
## 95% CI: [ 0.124 ,  0.223 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.553  vs  0.331
## Difference:  0.222
## Chi-squared:  46.585
## Degrees of freedom:  1
## P-value:  0.000000000008772
## 95% CI: [ 0.16 ,  0.285 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Sonnet-4
## ----------------------------------------
## Proportions:  0.553  vs  0.417
## Difference:  0.137
## Chi-squared:  29.312
## Degrees of freedom:  1
## P-value:  0.00000006161
## 95% CI: [ 0.087 ,  0.186 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Opus-4.1
## ----------------------------------------
## Proportions:  0.553  vs  0.529
## Difference:  0.024
## Chi-squared:  0.469
## Degrees of freedom:  1
## P-value:  0.4933
## 95% CI: [ -0.041 ,  0.089 ]
## Significant:  NO
##
##  o3-GPT-Image vs GPT-5
## ----------------------------------------
## Proportions:  0.553  vs  0.67
## Difference:  -0.116
## Chi-squared:  21.894
## Degrees of freedom:  1
## P-value:  0.000002882
## 95% CI: [ -0.164 ,  -0.068 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-4.1
## ----------------------------------------
## Proportions:  0.666  vs  0.425
## Difference:  0.241
## Chi-squared:  84.647
## Degrees of freedom:  1
## P-value:  0.0000000000000000000003567
## 95% CI: [ 0.19 ,  0.293 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.666  vs  0.383
## Difference:  0.283
## Chi-squared:  115.659
## Degrees of freedom:  1
## P-value:  0.00000000000000000000000000005644
## 95% CI: [ 0.232 ,  0.334 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.666  vs  0.42
## Difference:  0.247
## Chi-squared:  88.243
## Degrees of freedom:  1
## P-value:  0.00000000000000000000005789
## 95% CI: [ 0.195 ,  0.298 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o4-mini
## ----------------------------------------
## Proportions:  0.666  vs  0.53
## Difference:  0.136
## Chi-squared:  27.478
## Degrees of freedom:  1
## P-value:  0.0000001589
## 95% CI: [ 0.084 ,  0.188 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.666  vs  0.462
## Difference:  0.204
## Chi-squared:  61.125
## Degrees of freedom:  1
## P-value:  0.000000000000005357
## 95% CI: [ 0.153 ,  0.256 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.666  vs  0.38
## Difference:  0.286
## Chi-squared:  118.136
## Degrees of freedom:  1
## P-value:  0.000000000000000000000000001619
## 95% CI: [ 0.235 ,  0.337 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.666  vs  0.331
## Difference:  0.335
## Chi-squared:  102.513
## Degrees of freedom:  1
## P-value:  0.000000000000000000000004285
## 95% CI: [ 0.271 ,  0.399 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Sonnet-4
## ----------------------------------------
## Proportions:  0.666  vs  0.417
## Difference:  0.249
## Chi-squared:  90.129
## Degrees of freedom:  1
## P-value:  0.000000000000000000002231
## 95% CI: [ 0.198 ,  0.301 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Opus-4.1
## ----------------------------------------
## Proportions:  0.666  vs  0.529
## Difference:  0.137
## Chi-squared:  17.535
## Degrees of freedom:  1
## P-value:  0.00002821
## 95% CI: [ 0.07 ,  0.203 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-5
## ----------------------------------------
## Proportions:  0.666  vs  0.67
## Difference:  -0.003
## Chi-squared:  0.007
## Degrees of freedom:  1
## P-value:  0.9336
## 95% CI: [ -0.053 ,  0.047 ]
## Significant:  NO
##
##  GPT-4.1 vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.425  vs  0.383
## Difference:  0.042
## Chi-squared:  1.999
## Degrees of freedom:  1
## P-value:  0.1574
## 95% CI: [ -0.015 ,  0.099 ]
## Significant:  NO
##
##  GPT-4.1 vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.425  vs  0.42
## Difference:  0.005
## Chi-squared:  0.015
## Degrees of freedom:  1
## P-value:  0.9017
## 95% CI: [ -0.052 ,  0.063 ]
## Significant:  NO
##
##  GPT-4.1 vs o4-mini
## ----------------------------------------
## Proportions:  0.425  vs  0.53
## Difference:  -0.105
## Chi-squared:  12.951
## Degrees of freedom:  1
## P-value:  0.0003197
## 95% CI: [ -0.163 ,  -0.048 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.425  vs  0.462
## Difference:  -0.037
## Chi-squared:  1.519
## Degrees of freedom:  1
## P-value:  0.2177
## 95% CI: [ -0.095 ,  0.021 ]
## Significant:  NO
```

```
##
##  GPT-4.1 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.425  vs  0.38
## Difference:  0.045
## Chi-squared:  2.321
## Degrees of freedom:  1
## P-value:  0.1276
## 95% CI: [ -0.012 ,  0.102 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.425  vs  0.331
## Difference:  0.094
## Chi-squared:  6.957
## Degrees of freedom:  1
## P-value:  0.008347
## 95% CI: [ 0.025 ,  0.162 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs Sonnet-4
## ----------------------------------------
## Proportions:  0.425  vs  0.417
## Difference:  0.008
## Chi-squared:  0.047
## Degrees of freedom:  1
## P-value:  0.8277
## 95% CI: [ -0.05 ,  0.065 ]
## Significant:  NO
##
##  GPT-4.1 vs Opus-4.1
## ----------------------------------------
## Proportions:  0.425  vs  0.529
## Difference:  -0.105
## Chi-squared:  8.399
## Degrees of freedom:  1
## P-value:  0.003755
## 95% CI: [ -0.176 ,  -0.033 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs GPT-5
## ----------------------------------------
## Proportions:  0.425  vs  0.67
## Difference:  -0.245
## Chi-squared:  71.655
## Degrees of freedom:  1
## P-value:  0.00000000000000002564
## 95% CI: [ -0.301 ,  -0.189 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.383  vs  0.42
## Difference:  -0.037
## Chi-squared:  1.518
## Degrees of freedom:  1
## P-value:  0.2179
## 95% CI: [ -0.094 ,  0.021 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs o4-mini
## ----------------------------------------
## Proportions:  0.383  vs  0.53
## Difference:  -0.147
## Chi-squared:  25.599
## Degrees of freedom:  1
## P-value:  0.0000004203
## 95% CI: [ -0.205 ,  -0.09 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.383  vs  0.462
## Difference:  -0.079
## Chi-squared:  7.305
## Degrees of freedom:  1
## P-value:  0.006876
## 95% CI: [ -0.136 ,  -0.021 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.383  vs  0.38
## Difference:  0.003
## Chi-squared:  0.003
## Degrees of freedom:  1
## P-value:  0.9599
## 95% CI: [ -0.054 ,  0.06 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.383  vs  0.331
## Difference:  0.052
## Chi-squared:  2.104
## Degrees of freedom:  1
## P-value:  0.147
## 95% CI: [ -0.017 ,  0.12 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Sonnet-4
## ----------------------------------------
## Proportions:  0.383  vs  0.417
## Difference:  -0.034
## Chi-squared:  1.295
## Degrees of freedom:  1
## P-value:  0.2551
## 95% CI: [ -0.091 ,  0.023 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Opus-4.1
## ----------------------------------------
## Proportions:  0.383  vs  0.529
## Difference:  -0.146
## Chi-squared:  16.868
## Degrees of freedom:  1
## P-value:  0.00004007
## 95% CI: [ -0.217 ,  -0.075 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs GPT-5
## ----------------------------------------
## Proportions:  0.383  vs  0.67
## Difference:  -0.287
## Chi-squared:  97.737
## Degrees of freedom:  1
## P-value:  0.00000000000000000000004779
## 95% CI: [ -0.342 ,  -0.231 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs o4-mini
## ----------------------------------------
## Proportions:  0.42  vs  0.53
## Difference:  -0.111
## Chi-squared:  14.285
## Degrees of freedom:  1
## P-value:  0.0001571
## 95% CI: [ -0.168 ,  -0.053 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.42  vs  0.462
## Difference:  -0.042
## Chi-squared:  2.001
## Degrees of freedom:  1
## P-value:  0.1572
## 95% CI: [ -0.1 ,  0.016 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.42  vs  0.38
## Difference:  0.04
## Chi-squared:  1.8
## Degrees of freedom:  1
## P-value:  0.1797
## 95% CI: [ -0.017 ,  0.097 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.42  vs  0.331
## Difference:  0.088
## Chi-squared:  6.207
## Degrees of freedom:  1
## P-value:  0.01272
## 95% CI: [ 0.02 ,  0.157 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs Sonnet-4
## ----------------------------------------
## Proportions:  0.42  vs  0.417
## Difference:  0.003
## Chi-squared:  0.001
## Degrees of freedom:  1
## P-value:  0.9716
## 95% CI: [ -0.055 ,  0.06 ]
## Significant:  NO
##
##  ChatGPT-4o vs Opus-4.1
## ----------------------------------------
## Proportions:  0.42  vs  0.529
## Difference:  -0.11
## Chi-squared:  9.285
## Degrees of freedom:  1
## P-value:  0.002311
## 95% CI: [ -0.181 ,  -0.038 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs GPT-5
## ----------------------------------------
## Proportions:  0.42  vs  0.67
## Difference:  -0.25
## Chi-squared:  74.672
## Degrees of freedom:  1
## P-value:  0.000000000000000005558
## 95% CI: [ -0.306 ,  -0.194 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Gemini-2.5-Pro
## ----------------------------------------
## Proportions:  0.53  vs  0.462
## Difference:  0.068
## Chi-squared:  5.349
## Degrees of freedom:  1
## P-value:  0.02074
## 95% CI: [ 0.01 ,  0.127 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.53  vs  0.38
## Difference:  0.15
## Chi-squared:  26.707
## Degrees of freedom:  1
## P-value:  0.0000002367
## 95% CI: [ 0.093 ,  0.208 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.53  vs  0.331
## Difference:  0.199
## Chi-squared:  31.073
## Degrees of freedom:  1
## P-value:  0.00000002485
## 95% CI: [ 0.13 ,  0.268 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Sonnet-4
## ----------------------------------------
## Proportions:  0.53  vs  0.417
## Difference:  0.113
## Chi-squared:  15.001
## Degrees of freedom:  1
## P-value:  0.0001075
## 95% CI: [ 0.056 ,  0.171 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Opus-4.1
## ----------------------------------------
## Proportions:  0.53  vs  0.529
## Difference:  0.001
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.069 ,  0.071 ]
## Significant:  NO
##
##  o4-mini vs GPT-5
## ----------------------------------------
## Proportions:  0.53  vs  0.67
## Difference:  -0.139
## Chi-squared:  23.742
## Degrees of freedom:  1
## P-value:  0.000001102
## 95% CI: [ -0.196 ,  -0.083 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5-Pro vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.462  vs  0.38
## Difference:  0.082
## Chi-squared:  7.907
## Degrees of freedom:  1
## P-value:  0.004923
## 95% CI: [ 0.024 ,  0.139 ]
## Significant:  YES (p < 0.05)
```

```
##
##  Gemini-2.5-Pro vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.462  vs  0.331
## Difference:  0.131
## Chi-squared:  13.497
## Degrees of freedom:  1
## P-value:  0.000239
## 95% CI: [ 0.062 ,  0.2 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5-Pro vs Sonnet-4
## ----------------------------------------
## Proportions:  0.462  vs  0.417
## Difference:  0.045
## Chi-squared:  2.275
## Degrees of freedom:  1
## P-value:  0.1315
## 95% CI: [ -0.013 ,  0.103 ]
## Significant:  NO
##
##  Gemini-2.5-Pro vs Opus-4.1
## ----------------------------------------
## Proportions:  0.462  vs  0.529
## Difference:  -0.068
## Chi-squared:  3.394
## Degrees of freedom:  1
## P-value:  0.06541
## 95% CI: [ -0.139 ,  0.004 ]
## Significant:  NO
##
##  Gemini-2.5-Pro vs GPT-5
## ----------------------------------------
## Proportions:  0.462  vs  0.67
## Difference:  -0.208
## Chi-squared:  51.944
## Degrees of freedom:  1
## P-value:  0.0000000000005711
## 95% CI: [ -0.264 ,  -0.151 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs Gemini-2.0-Flash-Images
## ----------------------------------------
## Proportions:  0.38  vs  0.331
## Difference:  0.049
## Chi-squared:  1.854
## Degrees of freedom:  1
## P-value:  0.1733
## 95% CI: [ -0.02 ,  0.117 ]
## Significant:  NO
##
##  Gemini-2.0-Flash vs Sonnet-4
## ----------------------------------------
## Proportions:  0.38  vs  0.417
## Difference:  -0.037
## Chi-squared:  1.556
## Degrees of freedom:  1
## P-value:  0.2122
## 95% CI: [ -0.094 ,  0.02 ]
## Significant:  NO
##
##  Gemini-2.0-Flash vs Opus-4.1
## ---------------------------------------
## Proportions:  0.38  vs  0.529
## Difference:  -0.149
## Chi-squared:  17.617
## Degrees of freedom:  1
## P-value:  0.00002701
## 95% CI: [ -0.22 ,  -0.078 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs GPT-5
## ---------------------------------------
## Proportions:  0.38  vs  0.67
## Difference:  -0.29
## Chi-squared:  99.826
## Degrees of freedom:  1
## P-value:  0.00000000000000000000001664
## 95% CI: [ -0.345 ,  -0.234 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-Images vs Sonnet-4
## ---------------------------------------
## Proportions:  0.331  vs  0.417
## Difference:  -0.086
## Chi-squared:  5.836
## Degrees of freedom:  1
## P-value:  0.01571
## 95% CI: [ -0.155 ,  -0.017 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-Images vs Opus-4.1
## ---------------------------------------
## Proportions:  0.331  vs  0.529
## Difference:  -0.198
## Chi-squared:  23.247
## Degrees of freedom:  1
## P-value:  0.000001425
## 95% CI: [ -0.279 ,  -0.117 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-Images vs GPT-5
## ---------------------------------------
## Proportions:  0.331  vs  0.67
## Difference:  -0.339
## Chi-squared:  91.529
## Degrees of freedom:  1
## P-value:  0.0000000000000000000011
## 95% CI: [ -0.406 ,  -0.271 ]
## Significant:  YES (p < 0.05)
##
##  Sonnet-4 vs Opus-4.1
## ----------------------------------------
## Proportions:  0.417  vs  0.529
## Difference:  -0.112
## Chi-squared:  9.76
## Degrees of freedom:  1
## P-value:  0.001783
## 95% CI: [ -0.184 ,  -0.041 ]
## Significant:  YES (p < 0.05)
##
##  Sonnet-4 vs GPT-5
## ----------------------------------------
## Proportions:  0.417  vs  0.67
## Difference:  -0.253
## Chi-squared:  76.256
## Degrees of freedom:  1
## P-value:  0.00000000000000000002492
## 95% CI: [ -0.309 ,  -0.197 ]
## Significant:  YES (p < 0.05)
##
##  Opus-4.1 vs GPT-5
## ----------------------------------------
## Proportions:  0.529  vs  0.67
## Difference:  -0.14
## Chi-squared:  16.17
## Degrees of freedom:  1
## P-value:  0.0000579
## 95% CI: [ -0.211 ,  -0.07 ]
## 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  signifi
## ------------  ---------------------------------------------  -------  ------------  --------  -------
## X-squared     Humans vs o3                                    -0.094    28.6308876    0.0000  TRUE
## X-squared1    Humans vs o3-GPT-Image                          -0.006     0.1425741    0.7057  FALSE
## X-squared2    Humans vs o3-Pro                                -0.119    45.7603911    0.0000  TRUE
## X-squared3    Humans vs GPT-4.1                                0.122    32.9868982    0.0000  TRUE
## X-squared4    Humans vs GPT-4.1-GPT-Image                       0.164    59.4818080    0.0000  TRUE
## X-squared5    Humans vs ChatGPT-4o                              0.128    35.8604451    0.0000  TRUE
## X-squared6    Humans vs o4-mini                                 0.017     0.5765459    0.4477  FALSE
## X-squared7    Humans vs Gemini-2.5-Pro                          0.085    15.9603007    0.0001  TRUE
## X-squared8    Humans vs Gemini-2.0-Flash                        0.167    61.7405958    0.0000  TRUE
## X-squared9    Humans vs Gemini-2.0-Flash-Images                 0.216    53.2959199    0.0000  TRUE
## X-squared10   Humans vs Sonnet-4                                0.130    37.3919075    0.0000  TRUE
## X-squared11   Humans vs Opus-4.1                                0.018     0.2989942    0.5845  FALSE
## X-squared12   Humans vs GPT-5                                  -0.123    33.3019857    0.0000  TRUE
## X-squared13   o3 vs o3-GPT-Image                                0.088    16.1391378    0.0001  TRUE
## X-squared14   o3 vs o3-Pro                                     -0.025     1.1064986    0.2928  FALSE
## X-squared15   o3 vs GPT-4.1                                     0.217    67.6166340    0.0000  TRUE
## X-squared16   o3 vs GPT-4.1-GPT-Image                           0.258    95.7496135    0.0000  TRUE
## X-squared17   o3 vs ChatGPT-4o                                  0.222    70.8494853    0.0000  TRUE
## X-squared18   o3 vs o4-mini                                     0.111    18.0851641    0.0000  TRUE
## X-squared19   o3 vs Gemini-2.5-Pro                              0.180    46.7193652    0.0000  TRUE
## X-squared20   o3 vs Gemini-2.0-Flash                            0.262    98.0174945    0.0000  TRUE
## X-squared21   o3 vs Gemini-2.0-Flash-Images                     0.310    86.8942530    0.0000  TRUE
## X-squared22   o3 vs Sonnet-4                                    0.225    72.5490940    0.0000  TRUE
## X-squared23   o3 vs Opus-4.1                                    0.112    11.4662661    0.0007  TRUE
## X-squared24   o3 vs GPT-5                                      -0.028     1.1383107    0.2860  FALSE
## X-squared25   o3-GPT-Image vs o3-Pro                           -0.113    26.8195138    0.0000  TRUE
## X-squared26   o3-GPT-Image vs GPT-4.1                           0.129    26.0074801    0.0000  TRUE
## X-squared27   o3-GPT-Image vs GPT-4.1-GPT-Image                 0.170    45.8021597    0.0000  TRUE
## X-squared28   o3-GPT-Image vs ChatGPT-4o                        0.134    28.1640608    0.0000  TRUE
## X-squared29   o3-GPT-Image vs o4-mini                           0.023     0.7821301    0.3765  FALSE
## X-squared30   o3-GPT-Image vs Gemini-2.5-Pro                    0.092    13.1161642    0.0003  TRUE
## X-squared31   o3-GPT-Image vs Gemini-2.0-Flash                  0.174    47.4839020    0.0000  TRUE
## X-squared32   o3-GPT-Image vs Gemini-2.0-Flash-Images           0.222    46.5852344    0.0000  TRUE
## X-squared33   o3-GPT-Image vs Sonnet-4                          0.137    29.3120309    0.0000  TRUE
## X-squared34   o3-GPT-Image vs Opus-4.1                          0.024     0.4692957    0.4933  FALSE
## X-squared35   o3-GPT-Image vs GPT-5                            -0.116    21.8935396    0.0000  TRUE
## X-squared36   o3-Pro vs GPT-4.1                                 0.241    84.6471507    0.0000  TRUE
## X-squared37   o3-Pro vs GPT-4.1-GPT-Image                       0.283   115.6594014    0.0000  TRUE
## X-squared38   o3-Pro vs ChatGPT-4o                              0.247    88.2427557    0.0000  TRUE
## X-squared39   o3-Pro vs o4-mini                                 0.136    27.4784870    0.0000  TRUE
## X-squared40   o3-Pro vs Gemini-2.5-Pro                          0.204    61.1246804    0.0000  TRUE
## X-squared41   o3-Pro vs Gemini-2.0-Flash                        0.286   118.1359936    0.0000  TRUE
## X-squared42   o3-Pro vs Gemini-2.0-Flash-Images                 0.335   102.5130386    0.0000  TRUE
## X-squared43   o3-Pro vs Sonnet-4                                0.249    90.1292627    0.0000  TRUE
## X-squared44   o3-Pro vs Opus-4.1                                0.137    17.5347059    0.0000  TRUE
## X-squared45   o3-Pro vs GPT-5                                  -0.003     0.0069479    0.9336  FALSE
## X-squared46   GPT-4.1 vs GPT-4.1-GPT-Image                      0.042     1.9993745    0.1574  FALSE
## X-squared47   GPT-4.1 vs ChatGPT-4o                             0.005     0.0152534    0.9017  FALSE
## X-squared48   GPT-4.1 vs o4-mini                               -0.105    12.9510207    0.0003  TRUE
## X-squared49   GPT-4.1 vs Gemini-2.5-Pro                        -0.037     1.5193687    0.2177  FALSE
## X-squared50   GPT-4.1 vs Gemini-2.0-Flash                       0.045     2.3210330    0.1276  FALSE
## X-squared51   GPT-4.1 vs Gemini-2.0-Flash-Images                0.094     6.9574196    0.0083  TRUE
## X-squared52   GPT-4.1 vs Sonnet-4                               0.008     0.0473504    0.8277  FALSE
## X-squared53   GPT-4.1 vs Opus-4.1                              -0.105     8.3986450    0.0038  TRUE
## X-squared54   GPT-4.1 vs GPT-5                                 -0.245    71.6546450    0.0000  TRUE
## X-squared55   GPT-4.1-GPT-Image vs ChatGPT-4o                  -0.037     1.5181623    0.2179  FALSE
## X-squared56   GPT-4.1-GPT-Image vs o4-mini                     -0.147    25.5989635    0.0000  TRUE
## X-squared57   GPT-4.1-GPT-Image vs Gemini-2.5-Pro              -0.079     7.3049825    0.0069  TRUE
## X-squared58   GPT-4.1-GPT-Image vs Gemini-2.0-Flash             0.003     0.0025231    0.9599  FALSE
## X-squared59   GPT-4.1-GPT-Image vs Gemini-2.0-Flash-Images      0.052     2.1035348    0.1470  FALSE
## X-squared60   GPT-4.1-GPT-Image vs Sonnet-4                    -0.034     1.2951527    0.2551  FALSE
## X-squared61   GPT-4.1-GPT-Image vs Opus-4.1                    -0.146    16.8680920    0.0000  TRUE
## X-squared62   GPT-4.1-GPT-Image vs GPT-5                       -0.287    97.7365295    0.0000  TRUE
## X-squared63   ChatGPT-4o vs o4-mini                            -0.111    14.2854157    0.0002  TRUE
## X-squared64   ChatGPT-4o vs Gemini-2.5-Pro                     -0.042     2.0005049    0.1572  FALSE
## X-squared65   ChatGPT-4o vs Gemini-2.0-Flash                    0.040     1.8000628    0.1797  FALSE
## X-squared66   ChatGPT-4o vs Gemini-2.0-Flash-Images             0.088     6.2071015    0.0127  TRUE
## X-squared67   ChatGPT-4o vs Sonnet-4                            0.003     0.0012675    0.9716  FALSE
## X-squared68   ChatGPT-4o vs Opus-4.1                           -0.110     9.2845807    0.0023  TRUE
## X-squared69   ChatGPT-4o vs GPT-5                              -0.250    74.6719492    0.0000  TRUE
## X-squared70   o4-mini vs Gemini-2.5-Pro                         0.068     5.3489240    0.0207  TRUE
## X-squared71   o4-mini vs Gemini-2.0-Flash                       0.150    26.7071233    0.0000  TRUE
## X-squared72   o4-mini vs Gemini-2.0-Flash-Images                0.199    31.0733308    0.0000  TRUE
## X-squared73   o4-mini vs Sonnet-4                               0.113    15.0010285    0.0001  TRUE
## X-squared74   o4-mini vs Opus-4.1                               0.001     0.0000000    1.0000  FALSE
## X-squared75   o4-mini vs GPT-5                                 -0.139    23.7419837    0.0000  TRUE
## X-squared76   Gemini-2.5-Pro vs Gemini-2.0-Flash                0.082     7.9073611    0.0049  TRUE
## X-squared77   Gemini-2.5-Pro vs Gemini-2.0-Flash-Images         0.131    13.4965103    0.0002  TRUE
## X-squared78   Gemini-2.5-Pro vs Sonnet-4                        0.045     2.2752755    0.1315  FALSE
## X-squared79   Gemini-2.5-Pro vs Opus-4.1                       -0.068     3.3944881    0.0654  FALSE
## X-squared80   Gemini-2.5-Pro vs GPT-5                          -0.208    51.9440409    0.0000  TRUE
## X-squared81   Gemini-2.0-Flash vs Gemini-2.0-Flash-Images       0.049     1.8541845    0.1733  FALSE
## X-squared82   Gemini-2.0-Flash vs Sonnet-4                     -0.037     1.5564724    0.2122  FALSE
## X-squared83   Gemini-2.0-Flash vs Opus-4.1                     -0.149    17.6174219    0.0000  TRUE
## X-squared84   Gemini-2.0-Flash vs GPT-5                        -0.290    99.8257607    0.0000  TRUE
## X-squared85   Gemini-2.0-Flash-Images vs Sonnet-4              -0.086     5.8355995    0.0157  TRUE
## X-squared86   Gemini-2.0-Flash-Images vs Opus-4.1              -0.198    23.2466529    0.0000  TRUE
## X-squared87   Gemini-2.0-Flash-Images vs GPT-5                 -0.339    91.5289929    0.0000  TRUE
## X-squared88   Sonnet-4 vs Opus-4.1                             -0.112     9.7604857    0.0018  TRUE
## X-squared89   Sonnet-4 vs GPT-5                                -0.253    76.2556458    0.0000  TRUE
## X-squared90   Opus-4.1 vs GPT-5                                -0.140    16.1702630    0.0001  TRUE
```

```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: 66
```

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

```
##   Percentage significant: 72.5 %
```

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

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

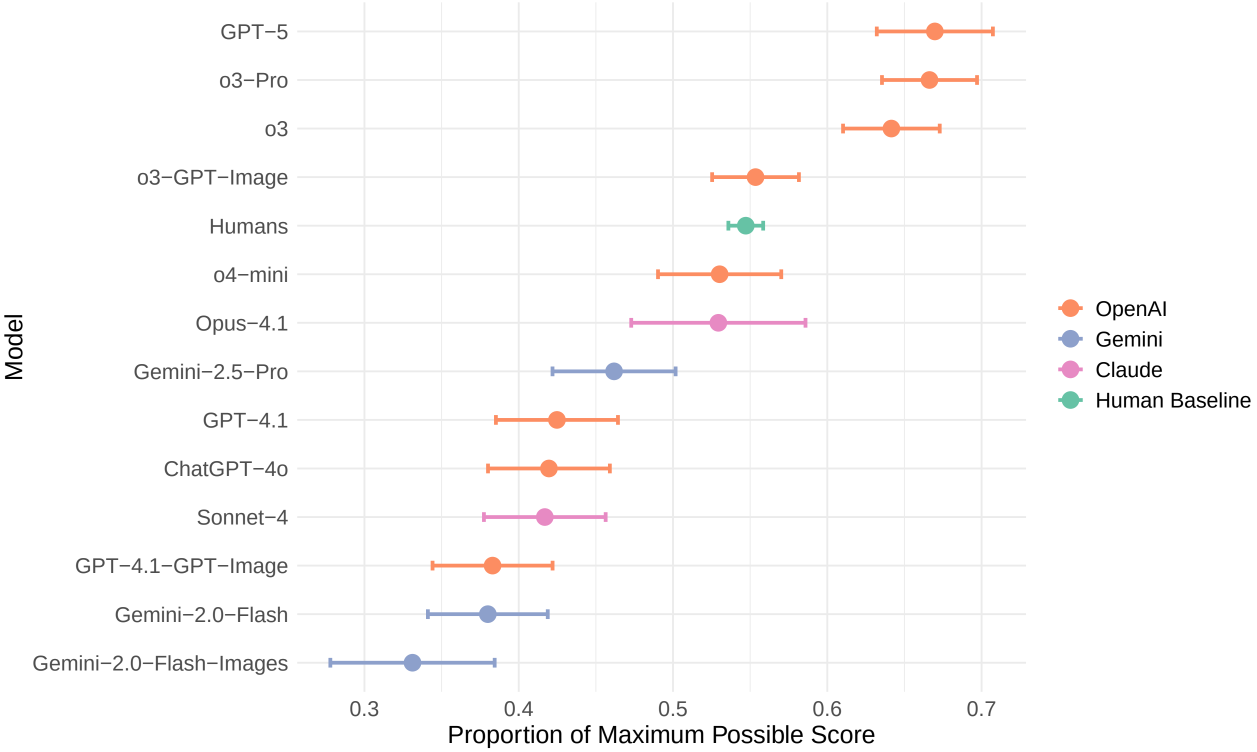
```r
collapsed_sig <- collapsed_results[collapsed_results$significant, c("comparison", "diff", "p_value")]
if (nrow(collapsed_sig) > 0) {
  print(kable(collapsed_sig, format = "simple", digits = 4))
} else {
  cat("  None\n")
}
```

```
##
##
##                comparison                                      diff    p_value
## ------------  ------------------------------------------  --------  --------
## X-squared     Humans vs o3                                 -0.0944    0.0000
## X-squared2    Humans vs o3-Pro                             -0.1191    0.0000
## X-squared3    Humans vs GPT-4.1                             0.1224    0.0000
## X-squared4    Humans vs GPT-4.1-GPT-Image                   0.1641    0.0000
## X-squared5    Humans vs ChatGPT-4o                          0.1276    0.0000
## X-squared7    Humans vs Gemini-2.5-Pro                      0.0854    0.0001
## X-squared8    Humans vs Gemini-2.0-Flash                    0.1672    0.0000
## X-squared9    Humans vs Gemini-2.0-Flash-Images             0.2160    0.0000
## X-squared10   Humans vs Sonnet-4                            0.1303    0.0000
## X-squared12   Humans vs GPT-5                              -0.1225    0.0000
## X-squared13   o3 vs o3-GPT-Image                            0.0881    0.0001
## X-squared15   o3 vs GPT-4.1                                 0.2168    0.0000
## X-squared16   o3 vs GPT-4.1-GPT-Image                       0.2585    0.0000
## X-squared17   o3 vs ChatGPT-4o                              0.2220    0.0000
## X-squared18   o3 vs o4-mini                                 0.1113    0.0000
## X-squared19   o3 vs Gemini-2.5-Pro                          0.1798    0.0000
## X-squared20   o3 vs Gemini-2.0-Flash                        0.2616    0.0000
## X-squared21   o3 vs Gemini-2.0-Flash-Images                 0.3104    0.0000
## X-squared22   o3 vs Sonnet-4                                0.2246    0.0000
## X-squared23   o3 vs Opus-4.1                                0.1121    0.0007
## X-squared25   o3-GPT-Image vs o3-Pro                       -0.1128    0.0000
## X-squared26   o3-GPT-Image vs GPT-4.1                       0.1287    0.0000
## X-squared27   o3-GPT-Image vs GPT-4.1-GPT-Image             0.1704    0.0000
## X-squared28   o3-GPT-Image vs ChatGPT-4o                    0.1339    0.0000
## X-squared30   o3-GPT-Image vs Gemini-2.5-Pro                0.0917    0.0003
## X-squared31   o3-GPT-Image vs Gemini-2.0-Flash              0.1735    0.0000
## X-squared32   o3-GPT-Image vs Gemini-2.0-Flash-Images       0.2223    0.0000
## X-squared33   o3-GPT-Image vs Sonnet-4                      0.1366    0.0000
## X-squared35   o3-GPT-Image vs GPT-5                        -0.1162    0.0000
## X-squared36   o3-Pro vs GPT-4.1                             0.2415    0.0000
## X-squared37   o3-Pro vs GPT-4.1-GPT-Image                   0.2832    0.0000
## X-squared38   o3-Pro vs ChatGPT-4o                          0.2467    0.0000
## X-squared39   o3-Pro vs o4-mini                             0.1360    0.0000
## X-squared40   o3-Pro vs Gemini-2.5-Pro                      0.2045    0.0000
## X-squared41   o3-Pro vs Gemini-2.0-Flash                    0.2863    0.0000
## X-squared42   o3-Pro vs Gemini-2.0-Flash-Images             0.3351    0.0000
## X-squared43   o3-Pro vs Sonnet-4                            0.2493    0.0000
## X-squared44   o3-Pro vs Opus-4.1                            0.1368    0.0000
## X-squared48   GPT-4.1 vs o4-mini                           -0.1054    0.0003
## X-squared51   GPT-4.1 vs Gemini-2.0-Flash-Images            0.0936    0.0083
## X-squared53   GPT-4.1 vs Opus-4.1                          -0.1046    0.0038
## X-squared54   GPT-4.1 vs GPT-5                             -0.2449    0.0000
## X-squared56   GPT-4.1-GPT-Image vs o4-mini                 -0.1472    0.0000
## X-squared57   GPT-4.1-GPT-Image vs Gemini-2.5-Pro          -0.0787    0.0069
## X-squared61   GPT-4.1-GPT-Image vs Opus-4.1                -0.1464    0.0000
## X-squared62   GPT-4.1-GPT-Image vs GPT-5                   -0.2867    0.0000
## X-squared63   ChatGPT-4o vs o4-mini                        -0.1106    0.0002
## X-squared66   ChatGPT-4o vs Gemini-2.0-Flash-Images         0.0884    0.0127
## X-squared68   ChatGPT-4o vs Opus-4.1                       -0.1098    0.0023
## X-squared69   ChatGPT-4o vs GPT-5                          -0.2501    0.0000
## X-squared70   o4-mini vs Gemini-2.5-Pro                     0.0684    0.0207
## X-squared71   o4-mini vs Gemini-2.0-Flash                   0.1502    0.0000
## X-squared72   o4-mini vs Gemini-2.0-Flash-Images            0.1991    0.0000
## X-squared73   o4-mini vs Sonnet-4                           0.1133    0.0001
## X-squared75   o4-mini vs GPT-5                             -0.1395    0.0000
## X-squared76   Gemini-2.5-Pro vs Gemini-2.0-Flash            0.0818    0.0049
## X-squared77   Gemini-2.5-Pro vs Gemini-2.0-Flash-Images     0.1306    0.0002
## X-squared80   Gemini-2.5-Pro vs GPT-5                      -0.2079    0.0000
## X-squared83   Gemini-2.0-Flash vs Opus-4.1                 -0.1494    0.0000
## X-squared84   Gemini-2.0-Flash vs GPT-5                    -0.2897    0.0000
## X-squared85   Gemini-2.0-Flash-Images vs Sonnet-4          -0.0858    0.0157
## X-squared86   Gemini-2.0-Flash-Images vs Opus-4.1          -0.1982    0.0000
## X-squared87   Gemini-2.0-Flash-Images vs GPT-5             -0.3386    0.0000
## X-squared88   Sonnet-4 vs Opus-4.1                         -0.1125    0.0018
## X-squared89   Sonnet-4 vs GPT-5                            -0.2528    0.0000
## X-squared90   Opus-4.1 vs GPT-5                            -0.1403    0.0001
```

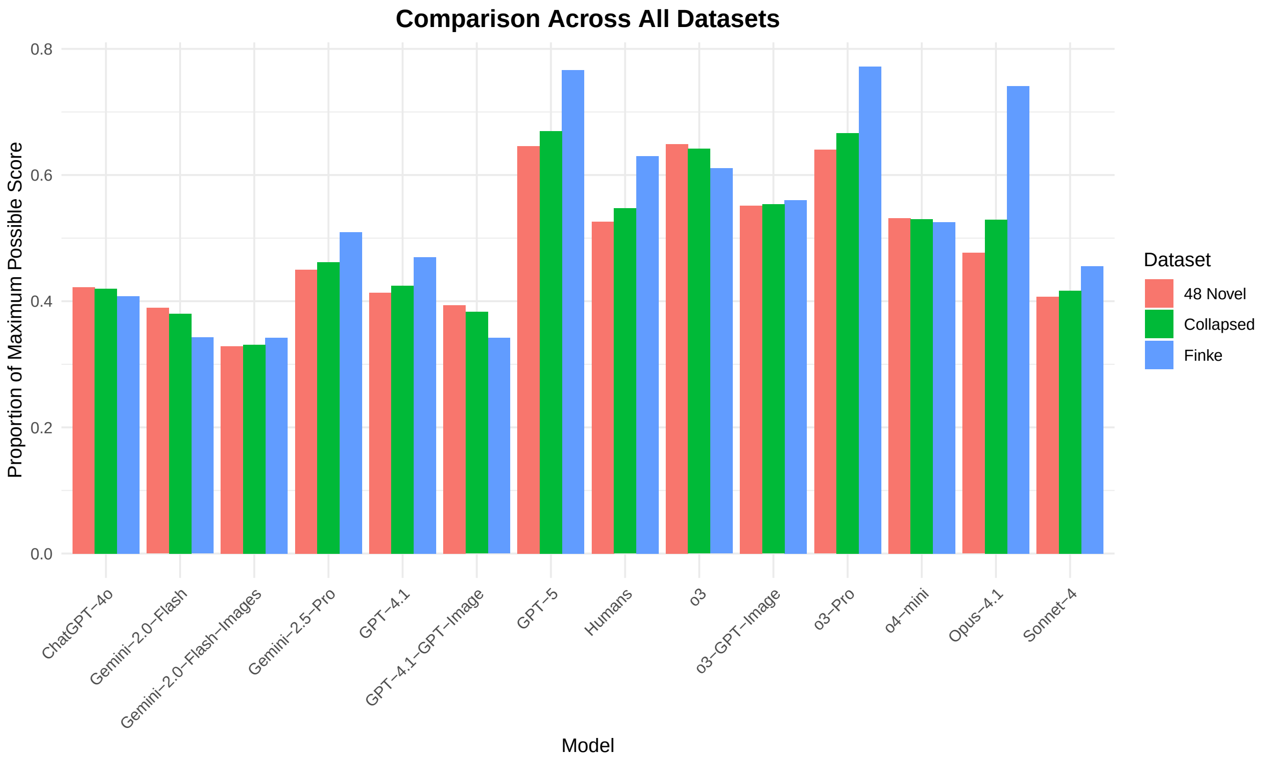
Visualization of Collapsed Data

```r
# Plot proportions with confidence intervals for collapsed data
collapsed_plot <- ggplot(collapsed_data, aes(x = reorder(model, proportion), y = proportion)) +
  geom_point(aes(color = color), size = 4) +
  geom_errorbar(aes(ymin = proportion - 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    ymax = proportion + 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    color = color),
                width = 0.2, size = 1) +
  coord_flip() +
  theme_minimal() +
  labs(
      x = "Model",
      y = "Proportion of Maximum Possible Score") +
  theme(plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
       axis.text = element_text(size = 12),
       axis.title = element_text(size = 14),
       legend.text = element_text(size = 12)) +
  scale_color_manual(
    values = c("#fc8d62", "#8da0cb", "#e78ac3", "#66c2a5"),
    name = element_blank(),
    breaks = c("#fc8d62", "#8da0cb", "#e78ac3", "#66c2a5"),
    labels = c("OpenAI", "Gemini", "Claude", "Human Baseline")
  )

# red #66c2a5
# blue #fc8d62
# green #8da0cb
# purple #e78ac3
```

```
print(collapsed_plot)
```



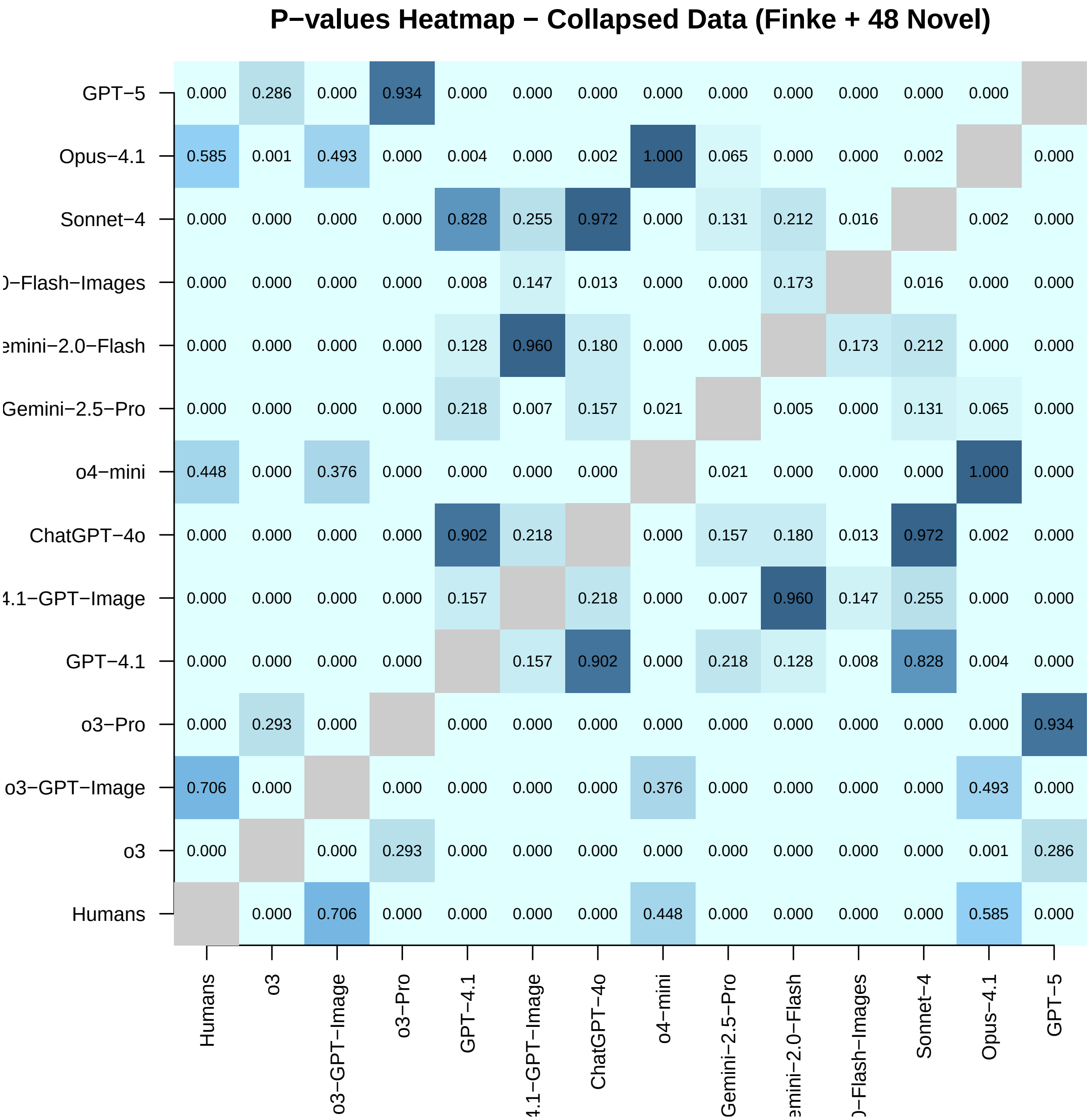
```
# Create a comparison plot showing all three datasets

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 of Maximum Possible Score",
       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

```r
# 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)
```

```r
# 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−Pro	emini−2.0−Flash	)−Flash−Images	Sonnet−4	Opus−4.1	GPT−5
GPT−5	0.000	0.286	0.000	0.934	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
Opus−4.1	0.585	0.001	0.493	0.000	0.004	0.000	0.002	1.000	0.065	0.000	0.000	0.002		0.000
Sonnet−4	0.000	0.000	0.000	0.000	0.828	0.255	0.972	0.000	0.131	0.212	0.016		0.002	0.000
)−Flash−Images	0.000	0.000	0.000	0.000	0.008	0.147	0.013	0.000	0.000	0.173		0.016	0.000	0.000
emini−2.0−Flash	0.000	0.000	0.000	0.000	0.128	0.960	0.180	0.000	0.005		0.173	0.212	0.000	0.000
Gemini−2.5−Pro	0.000	0.000	0.000	0.000	0.218	0.007	0.157	0.021		0.005	0.000	0.131	0.065	0.000
o4−mini	0.448	0.000	0.376	0.000	0.000	0.000	0.000		0.021	0.000	0.000	0.000	1.000	0.000
ChatGPT−4o	0.000	0.000	0.000	0.000	0.902	0.218		0.000	0.157	0.180	0.013	0.972	0.002	0.000
4.1−GPT−Image	0.000	0.000	0.000	0.000	0.157		0.218	0.000	0.007	0.960	0.147	0.255	0.000	0.000
GPT−4.1	0.000	0.000	0.000	0.000		0.157	0.902	0.000	0.218	0.128	0.008	0.828	0.004	0.000
o3−Pro	0.000	0.293	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.934
o3−GPT−Image	0.706	0.000		0.000	0.000	0.000	0.000	0.376	0.000	0.000	0.000	0.000	0.493	0.000
o3	0.000		0.000	0.293	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.286
Humans		0.000	0.706	0.000	0.000	0.000	0.000	0.448	0.000	0.000	0.000	0.000	0.585	0.000

# Reasoning Variation Analysis

Finke

```r
# Test all combinations for Finke reasoning variations
finke_reasoning_results <- test_all_combinations(finke_reasoning_data, "Finke Reasoning Variations")
# Display results
cat("All Pairwise Comparisons for Finke Reasoning Variations:\n")
```

```
## All Pairwise Comparisons for Finke Reasoning Variations:
```

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

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

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

```
##
##  Humans vs o3-High
## ----------------------------------------
## Proportions:  0.63  vs  0.611
## Difference:  0.02
## Chi-squared:  0.189
## Degrees of freedom:  1
## P-value:  0.6636
## 95% CI: [ -0.059 ,  0.098 ]
## Significant:  NO
##
##  Humans vs o3-Medium
## ----------------------------------------
## Proportions:  0.63  vs  0.574
## Difference:  0.057
## Chi-squared:  0.568
## Degrees of freedom:  1
## P-value:  0.4509
## 95% CI: [ -0.08 ,  0.193 ]
## Significant:  NO
##
##  Humans vs o3-Low
## ----------------------------------------
## Proportions:  0.63  vs  0.623
## Difference:  0.007
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.125 ,  0.139 ]
## Significant:  NO
##
##  Humans vs GPT-5-High
## ----------------------------------------
## Proportions:  0.63  vs  0.766
## Difference:  -0.136
## Chi-squared:  8.354
## Degrees of freedom:  1
## P-value:  0.003847
## 95% CI: [ -0.22 ,  -0.052 ]
## Significant:  YES (p < 0.05)
```

```
##
##  Humans vs o3-Pro
## ---------------------------------------
## Proportions:  0.63  vs  0.772
## Difference:  -0.141
## Chi-squared:  13.467
## Degrees of freedom:  1
## P-value:  0.0002428
## 95% CI: [ -0.211 ,  -0.072 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-5-Medium
## ---------------------------------------
## Proportions:  0.63  vs  0.633
## Difference:  -0.003
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.131 ,  0.124 ]
## Significant:  NO
##
##  Humans vs GPT-5-Low
## ---------------------------------------
## Proportions:  0.63  vs  0.556
## Difference:  0.074
## Chi-squared:  1.063
## Degrees of freedom:  1
## P-value:  0.3026
## 95% CI: [ -0.062 ,  0.211 ]
## Significant:  NO
##
##  Humans vs GPT-5-Minimal
## ---------------------------------------
## Proportions:  0.63  vs  0.366
## Difference:  0.265
## Chi-squared:  16.06
## Degrees of freedom:  1
## P-value:  0.00006135
## 95% CI: [ 0.132 ,  0.398 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o4-mini-High
## ---------------------------------------
## Proportions:  0.63  vs  0.525
## Difference:  0.105
## Chi-squared:  4.797
## Degrees of freedom:  1
## P-value:  0.0285
## 95% CI: [ 0.008 ,  0.202 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.63  vs  0.467
## Difference:  0.163
## Chi-squared:  11.896
## Degrees of freedom:  1
## P-value:  0.0005627
## 95% CI: [ 0.066 ,  0.26 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.63  vs  0.56
## Difference:  0.07
## Chi-squared:  4.009
## Degrees of freedom:  1
## P-value:  0.04525
## 95% CI: [ 0 ,  0.14 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.63  vs  0.506
## Difference:  0.125
## Chi-squared:  3.312
## Degrees of freedom:  1
## P-value:  0.06877
## 95% CI: [ -0.013 ,  0.262 ]
## Significant:  NO
##
##  o3-High vs o3-Medium
## ----------------------------------------
## Proportions:  0.611  vs  0.574
## Difference:  0.037
## Chi-squared:  0.125
## Degrees of freedom:  1
## P-value:  0.7234
## 95% CI: [ -0.118 ,  0.192 ]
## Significant:  NO
##
##  o3-High vs o3-Low
## ----------------------------------------
## Proportions:  0.611  vs  0.623
## Difference:  -0.012
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  0.9847
## 95% CI: [ -0.165 ,  0.14 ]
## Significant:  NO
##
##  o3-High vs GPT-5-High
## ----------------------------------------
## Proportions:  0.611  vs  0.766
## Difference:  -0.156
## Chi-squared:  7.237
## Degrees of freedom:  1
## P-value:  0.007141
## 95% CI: [ -0.267 ,  -0.045 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o3-Pro
## ----------------------------------------
## Proportions:  0.611  vs  0.772
## Difference:  -0.161
## Chi-squared:  10.208
## Degrees of freedom:  1
## P-value:  0.001398
## 95% CI: [ -0.261 ,  -0.062 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.611  vs  0.633
## Difference:  -0.023
## Chi-squared:  0.027
## Degrees of freedom:  1
## P-value:  0.8706
## 95% CI: [ -0.175 ,  0.129 ]
## Significant:  NO
##
##  o3-High vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.611  vs  0.556
## Difference:  0.055
## Chi-squared:  0.353
## Degrees of freedom:  1
## P-value:  0.5522
## 95% CI: [ -0.101 ,  0.21 ]
## Significant:  NO
##
##  o3-High vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.611  vs  0.366
## Difference:  0.245
## Chi-squared:  9.942
## Degrees of freedom:  1
## P-value:  0.001616
## 95% CI: [ 0.093 ,  0.397 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o4-mini-High
## ----------------------------------------
## Proportions:  0.611  vs  0.525
## Difference:  0.085
## Chi-squared:  1.819
## Degrees of freedom:  1
## P-value:  0.1774
## 95% CI: [ -0.036 ,  0.207 ]
## Significant:  NO
##
##  o3-High vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.611  vs  0.467
## Difference:  0.144
## Chi-squared:  5.446
## Degrees of freedom:  1
## P-value:  0.01961
## 95% CI: [ 0.023 ,  0.265 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.611  vs  0.56
## Difference:  0.05
## Chi-squared:  0.869
## Degrees of freedom:  1
## P-value:  0.3511
## 95% CI: [ -0.05 ,  0.15 ]
## Significant:  NO
##
##  o3-High vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.611  vs  0.506
## Difference:  0.105
## Chi-squared:  1.63
## Degrees of freedom:  1
## P-value:  0.2016
## 95% CI: [ -0.051 ,  0.261 ]
## Significant:  NO
##
##  o3-Medium vs o3-Low
## ----------------------------------------
## Proportions:  0.574  vs  0.623
## Difference:  -0.049
## Chi-squared:  0.134
## Degrees of freedom:  1
## P-value:  0.7142
## 95% CI: [ -0.241 ,  0.142 ]
## Significant:  NO
##
##  o3-Medium vs GPT-5-High
## ----------------------------------------
## Proportions:  0.574  vs  0.766
## Difference:  -0.193
## Chi-squared:  6.205
## Degrees of freedom:  1
## P-value:  0.01274
## 95% CI: [ -0.351 ,  -0.034 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs o3-Pro
## ----------------------------------------
## Proportions:  0.574  vs  0.772
## Difference:  -0.198
## Chi-squared:  7.842
## Degrees of freedom:  1
## P-value:  0.005103
## 95% CI: [ -0.349 ,  -0.048 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.574  vs  0.633
## Difference:  -0.06
## Chi-squared:  0.234
## Degrees of freedom:  1
## P-value:  0.6286
## 95% CI: [ -0.251 ,  0.132 ]
## Significant:  NO
##
##  o3-Medium vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.574  vs  0.556
## Difference:  0.018
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  0.9914
## 95% CI: [ -0.176 ,  0.212 ]
## Significant:  NO
##
##  o3-Medium vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.574  vs  0.366
## Difference:  0.208
## Chi-squared:  4.411
## Degrees of freedom:  1
## P-value:  0.03571
## 95% CI: [ 0.017 ,  0.399 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs o4-mini-High
## ----------------------------------------
## Proportions:  0.574  vs  0.525
## Difference:  0.049
## Chi-squared:  0.209
## Degrees of freedom:  1
## P-value:  0.6473
## 95% CI: [ -0.118 ,  0.215 ]
## Significant:  NO
##
##  o3-Medium vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.574  vs  0.467
## Difference:  0.107
## Chi-squared:  1.421
## Degrees of freedom:  1
## P-value:  0.2333
## 95% CI: [ -0.059 ,  0.273 ]
## Significant:  NO
```

```
##
##  o3-Medium vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.574  vs  0.56
## Difference:  0.013
## Chi-squared:  0.002
## Degrees of freedom:  1
## P-value:  0.9683
## 95% CI: [ -0.137 ,  0.164 ]
## Significant:  NO
##
##  o3-Medium vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.574  vs  0.506
## Difference:  0.068
## Chi-squared:  0.318
## Degrees of freedom:  1
## P-value:  0.5729
## 95% CI: [ -0.127 ,  0.263 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-High
## ----------------------------------------
## Proportions:  0.623  vs  0.766
## Difference:  -0.143
## Chi-squared:  3.378
## Degrees of freedom:  1
## P-value:  0.06606
## 95% CI: [ -0.3 ,  0.013 ]
## Significant:  NO
##
##  o3-Low vs o3-Pro
## ----------------------------------------
## Proportions:  0.623  vs  0.772
## Difference:  -0.149
## Chi-squared:  4.366
## Degrees of freedom:  1
## P-value:  0.03666
## 95% CI: [ -0.297 ,  0 ]
## Significant:  YES (p < 0.05)
##
##  o3-Low vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.623  vs  0.633
## Difference:  -0.01
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.194 ,  0.173 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.623  vs  0.556
## Difference:  0.067
## Chi-squared:  0.315
## Degrees of freedom:  1
## P-value:  0.5746
## 95% CI: [ -0.125 ,  0.259 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.623  vs  0.366
## Difference:  0.257
## Chi-squared:  6.96
## Degrees of freedom:  1
## P-value:  0.008336
## 95% CI: [ 0.068 ,  0.447 ]
## Significant:  YES (p < 0.05)
##
##  o3-Low vs o4-mini-High
## ----------------------------------------
## Proportions:  0.623  vs  0.525
## Difference:  0.098
## Chi-squared:  1.185
## Degrees of freedom:  1
## P-value:  0.2763
## 95% CI: [ -0.066 ,  0.262 ]
## Significant:  NO
##
##  o3-Low vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.623  vs  0.467
## Difference:  0.156
## Chi-squared:  3.308
## Degrees of freedom:  1
## P-value:  0.06895
## 95% CI: [ -0.008 ,  0.32 ]
## Significant:  NO
##
##  o3-Low vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.623  vs  0.56
## Difference:  0.063
## Chi-squared:  0.536
## Degrees of freedom:  1
## P-value:  0.4639
## 95% CI: [ -0.085 ,  0.211 ]
## Significant:  NO
##
##  o3-Low vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.623  vs  0.506
## Difference:  0.117
## Chi-squared:  1.239
## Degrees of freedom:  1
## P-value:  0.2657
## 95% CI: [ -0.075 ,  0.31 ]
## Significant:  NO
##
##  GPT-5-High vs o3-Pro
## ----------------------------------------
## Proportions:  0.766  vs  0.772
## Difference:  -0.005
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.108 ,  0.097 ]
## Significant:  NO
##
##  GPT-5-High vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.766  vs  0.633
## Difference:  0.133
## Chi-squared:  2.883
## Degrees of freedom:  1
## P-value:  0.08952
## 95% CI: [ -0.023 ,  0.289 ]
## Significant:  NO
##
##  GPT-5-High vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.766  vs  0.556
## Difference:  0.21
## Chi-squared:  7.397
## Degrees of freedom:  1
## P-value:  0.006533
## 95% CI: [ 0.051 ,  0.37 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.766  vs  0.366
## Difference:  0.401
## Chi-squared:  25.935
## Degrees of freedom:  1
## P-value:  0.0000003531
## 95% CI: [ 0.245 ,  0.557 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o4-mini-High
## ----------------------------------------
## Proportions:  0.766  vs  0.525
## Difference:  0.241
## Chi-squared:  14.219
## Degrees of freedom:  1
## P-value:  0.0001627
## 95% CI: [ 0.116 ,  0.367 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.766  vs  0.467
## Difference:  0.299
## Chi-squared:  21.498
## Degrees of freedom:  1
## P-value:  0.000003543
## 95% CI: [ 0.174 ,  0.425 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.766  vs  0.56
## Difference:  0.206
## Chi-squared:  13.665
## Degrees of freedom:  1
## P-value:  0.0002185
## 95% CI: [ 0.101 ,  0.311 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.766  vs  0.506
## Difference:  0.261
## Chi-squared:  11.305
## Degrees of freedom:  1
## P-value:  0.0007731
## 95% CI: [ 0.101 ,  0.421 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.772  vs  0.633
## Difference:  0.138
## Chi-squared:  3.754
## Degrees of freedom:  1
## P-value:  0.05269
## 95% CI: [ -0.009 ,  0.286 ]
## Significant:  NO
##
##  o3-Pro vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.772  vs  0.556
## Difference:  0.216
## Chi-squared:  9.302
## Degrees of freedom:  1
## P-value:  0.002289
## 95% CI: [ 0.065 ,  0.367 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.772  vs  0.366
## Difference:  0.406
## Chi-squared:  31.768
## Degrees of freedom:  1
## P-value:  0.00000001737
## 95% CI: [ 0.259 ,  0.554 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o4-mini-High
## ----------------------------------------
## Proportions:  0.772  vs  0.525
## Difference:  0.247
## Chi-squared:  18.799
## Degrees of freedom:  1
## P-value:  0.00001452
## 95% CI: [ 0.131 ,  0.362 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.772  vs  0.467
## Difference:  0.305
## Chi-squared:  28.077
## Degrees of freedom:  1
## P-value:  0.0000001166
## 95% CI: [ 0.19 ,  0.42 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.772  vs  0.56
## Difference:  0.211
## Chi-squared:  19.304
## Degrees of freedom:  1
## P-value:  0.00001115
## 95% CI: [ 0.119 ,  0.304 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.772  vs  0.506
## Difference:  0.266
## Chi-squared:  14.071
## Degrees of freedom:  1
## P-value:  0.000176
## 95% CI: [ 0.114 ,  0.418 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.633  vs  0.556
## Difference:  0.078
## Chi-squared:  0.461
## Degrees of freedom:  1
## P-value:  0.4973
## 95% CI: [ -0.114 ,  0.269 ]
## Significant:  NO
```

```
##
##  GPT-5-Medium vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.633  vs  0.366
## Difference:  0.268
## Chi-squared:  7.574
## Degrees of freedom:  1
## P-value:  0.005922
## 95% CI: [ 0.079 ,  0.457 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs o4-mini-High
## ----------------------------------------
## Proportions:  0.633  vs  0.525
## Difference:  0.108
## Chi-squared:  1.494
## Degrees of freedom:  1
## P-value:  0.2216
## 95% CI: [ -0.055 ,  0.272 ]
## Significant:  NO
##
##  GPT-5-Medium vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.633  vs  0.467
## Difference:  0.167
## Chi-squared:  3.807
## Degrees of freedom:  1
## P-value:  0.05104
## 95% CI: [ 0.003 ,  0.33 ]
## Significant:  NO
##
##  GPT-5-Medium vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.633  vs  0.56
## Difference:  0.073
## Chi-squared:  0.772
## Degrees of freedom:  1
## P-value:  0.3795
## 95% CI: [ -0.074 ,  0.221 ]
## Significant:  NO
##
##  GPT-5-Medium vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.633  vs  0.506
## Difference:  0.128
## Chi-squared:  1.512
## Degrees of freedom:  1
## P-value:  0.2188
## 95% CI: [ -0.065 ,  0.32 ]
## Significant:  NO
##
##  GPT-5-Low vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.556  vs  0.366
## Difference:  0.19
## Chi-squared:  3.644
## Degrees of freedom:  1
## P-value:  0.05626
## 95% CI: [ -0.001 ,  0.382 ]
## Significant:  NO
##
##  GPT-5-Low vs o4-mini-High
## ----------------------------------------
## Proportions:  0.556  vs  0.525
## Difference:  0.031
## Chi-squared:  0.054
## Degrees of freedom:  1
## P-value:  0.8155
## 95% CI: [ -0.136 ,  0.198 ]
## Significant:  NO
##
##  GPT-5-Low vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.556  vs  0.467
## Difference:  0.089
## Chi-squared:  0.939
## Degrees of freedom:  1
## P-value:  0.3326
## 95% CI: [ -0.078 ,  0.256 ]
## Significant:  NO
##
##  GPT-5-Low vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.556  vs  0.56
## Difference:  -0.004
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.149 ,  0.141 ]
## Significant:  NO
##
##  GPT-5-Low vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.556  vs  0.506
## Difference:  0.05
## Chi-squared:  0.137
## Degrees of freedom:  1
## P-value:  0.7117
## 95% CI: [ -0.145 ,  0.245 ]
## Significant:  NO
##
##  GPT-5-Minimal vs o4-mini-High
## ----------------------------------------
## Proportions:  0.366  vs  0.525
## Difference:  -0.159
## Chi-squared:  3.468
## Degrees of freedom:  1
## P-value:  0.06256
## 95% CI: [ -0.323 ,  0.004 ]
## Significant:  NO
##
##  GPT-5-Minimal vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.366  vs  0.467
## Difference:  -0.101
## Chi-squared:  1.285
## Degrees of freedom:  1
## P-value:  0.257
## 95% CI: [ -0.265 ,  0.062 ]
## Significant:  NO
##
##  GPT-5-Minimal vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.366  vs  0.56
## Difference:  -0.195
## Chi-squared:  6.537
## Degrees of freedom:  1
## P-value:  0.01056
## 95% CI: [ -0.342 ,  -0.047 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Minimal vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.366  vs  0.506
## Difference:  -0.14
## Chi-squared:  1.858
## Degrees of freedom:  1
## P-value:  0.1729
## 95% CI: [ -0.332 ,  0.052 ]
## Significant:  NO
##
##  o4-mini-High vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.525  vs  0.467
## Difference:  0.058
## Chi-squared:  0.597
## Degrees of freedom:  1
## P-value:  0.4398
## 95% CI: [ -0.076 ,  0.193 ]
## Significant:  NO
##
##  o4-mini-High vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.525  vs  0.56
## Difference:  -0.035
## Chi-squared:  0.272
## Degrees of freedom:  1
## P-value:  0.6019
## 95% CI: [ -0.151 ,  0.08 ]
## Significant:  NO
##
##  o4-mini-High vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.525  vs  0.506
## Difference:  0.019
## Chi-squared:  0.008
## Degrees of freedom:  1
## P-value:  0.9301
## 95% CI: [ -0.148 ,  0.187 ]
## Significant:  NO
##
##  o4-mini-Medium vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.467  vs  0.56
## Difference:  -0.093
## Chi-squared:  2.443
## Degrees of freedom:  1
## P-value:  0.1181
## 95% CI: [ -0.209 ,  0.022 ]
## Significant:  NO
##
##  o4-mini-Medium vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.467  vs  0.506
## Difference:  -0.039
## Chi-squared:  0.111
## Degrees of freedom:  1
## P-value:  0.7396
## 95% CI: [ -0.206 ,  0.129 ]
## Significant:  NO
##
##  o3-GPT-Image-High vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.56  vs  0.506
## Difference:  0.055
## Chi-squared:  0.38
## Degrees of freedom:  1
## P-value:  0.5374
## 95% CI: [ -0.097 ,  0.206 ]
## Significant:  NO
```

```
# Summary table
finke_reasoning_summary <- finke_reasoning_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 Reasoning Variations:\n")
```

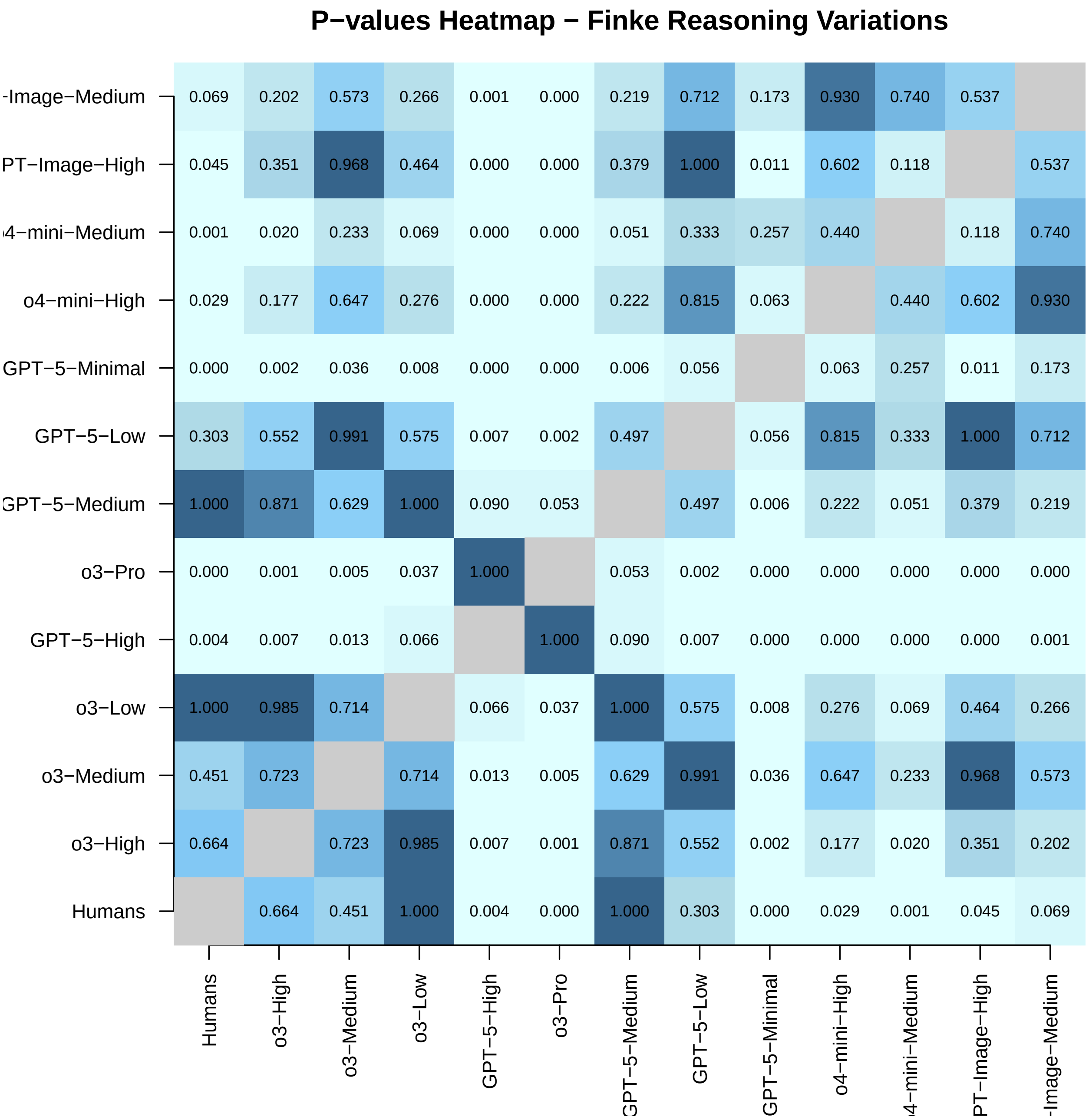
```
##
##
## Summary Table - Finke Reasoning Variations:
```

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

```
##
##
##                comparison                                  diff   chi_squared   p_value   significan
## ------------  -----------------------------------------  -------  ------------  --------  -----------
## X-squared     Humans vs o3-High                            0.020     0.1891561    0.6636  FALSE
## X-squared1    Humans vs o3-Medium                          0.057     0.5683218    0.4509  FALSE
## X-squared2    Humans vs o3-Low                             0.007     0.0000000    1.0000  FALSE
## X-squared3    Humans vs GPT-5-High                        -0.136     8.3544955    0.0038  TRUE
## X-squared4    Humans vs o3-Pro                            -0.141    13.4669053    0.0002  TRUE
## X-squared5    Humans vs GPT-5-Medium                      -0.003     0.0000000    1.0000  FALSE
## X-squared6    Humans vs GPT-5-Low                          0.074     1.0625196    0.3026  FALSE
## X-squared7    Humans vs GPT-5-Minimal                      0.265    16.0604213    0.0001  TRUE
## X-squared8    Humans vs o4-mini-High                       0.105     4.7972861    0.0285  TRUE
## X-squared9    Humans vs o4-mini-Medium                     0.163    11.8955544    0.0006  TRUE
## X-squared10   Humans vs o3-GPT-Image-High                  0.070     4.0094812    0.0452  TRUE
## X-squared11   Humans vs o3-GPT-Image-Medium                0.125     3.3122048    0.0688  FALSE
## X-squared12   o3-High vs o3-Medium                         0.037     0.1252689    0.7234  FALSE
## X-squared13   o3-High vs o3-Low                           -0.012     0.0003661    0.9847  FALSE
## X-squared14   o3-High vs GPT-5-High                       -0.156     7.2371887    0.0071  TRUE
## X-squared15   o3-High vs o3-Pro                           -0.161    10.2079814    0.0014  TRUE
## X-squared16   o3-High vs GPT-5-Medium                     -0.023     0.0265213    0.8706  FALSE
## X-squared17   o3-High vs GPT-5-Low                         0.055     0.3534477    0.5522  FALSE
## X-squared18   o3-High vs GPT-5-Minimal                     0.245     9.9417462    0.0016  TRUE
## X-squared19   o3-High vs o4-mini-High                      0.085     1.8190399    0.1774  FALSE
## X-squared20   o3-High vs o4-mini-Medium                    0.144     5.4460525    0.0196  TRUE
## X-squared21   o3-High vs o3-GPT-Image-High                 0.050     0.8693549    0.3511  FALSE
## X-squared22   o3-High vs o3-GPT-Image-Medium               0.105     1.6304149    0.2016  FALSE
## X-squared23   o3-Medium vs o3-Low                         -0.049     0.1341131    0.7142  FALSE
## X-squared24   o3-Medium vs GPT-5-High                     -0.193     6.2051551    0.0127  TRUE
## X-squared25   o3-Medium vs o3-Pro                         -0.198     7.8424357    0.0051  TRUE
## X-squared26   o3-Medium vs GPT-5-Medium                   -0.060     0.2339226    0.6286  FALSE
## X-squared27   o3-Medium vs GPT-5-Low                       0.018     0.0001154    0.9914  FALSE
## X-squared28   o3-Medium vs GPT-5-Minimal                   0.208     4.4109665    0.0357  TRUE
## X-squared29   o3-Medium vs o4-mini-High                    0.049     0.2092646    0.6473  FALSE
## X-squared30   o3-Medium vs o4-mini-Medium                  0.107     1.4209054    0.2333  FALSE
## X-squared31   o3-Medium vs o3-GPT-Image-High               0.013     0.0015816    0.9683  FALSE
## X-squared32   o3-Medium vs o3-GPT-Image-Medium             0.068     0.3179166    0.5729  FALSE
## X-squared33   o3-Low vs GPT-5-High                        -0.143     3.3782085    0.0661  FALSE
## X-squared34   o3-Low vs o3-Pro                            -0.149     4.3662115    0.0367  TRUE
## X-squared35   o3-Low vs GPT-5-Medium                      -0.010     0.0000000    1.0000  FALSE
## X-squared36   o3-Low vs GPT-5-Low                          0.067     0.3151208    0.5746  FALSE
## X-squared37   o3-Low vs GPT-5-Minimal                      0.257     6.9598362    0.0083  TRUE
## X-squared38   o3-Low vs o4-mini-High                       0.098     1.1850589    0.2763  FALSE
## X-squared39   o3-Low vs o4-mini-Medium                     0.156     3.3078003    0.0690  FALSE
## X-squared40   o3-Low vs o3-GPT-Image-High                  0.063     0.5364076    0.4639  FALSE
## X-squared41   o3-Low vs o3-GPT-Image-Medium                0.117     1.2386775    0.2657  FALSE
## X-squared42   GPT-5-High vs o3-Pro                        -0.005     0.0000000    1.0000  FALSE
## X-squared43   GPT-5-High vs GPT-5-Medium                   0.133     2.8829685    0.0895  FALSE
## X-squared44   GPT-5-High vs GPT-5-Low                      0.210     7.3970555    0.0065  TRUE
## X-squared45   GPT-5-High vs GPT-5-Minimal                  0.401    25.9353072    0.0000  TRUE
## X-squared46   GPT-5-High vs o4-mini-High                   0.241    14.2190034    0.0002  TRUE
## X-squared47   GPT-5-High vs o4-mini-Medium                 0.299    21.4975960    0.0000  TRUE
## X-squared48   GPT-5-High vs o3-GPT-Image-High              0.206    13.6649251    0.0002  TRUE
## X-squared49   GPT-5-High vs o3-GPT-Image-Medium            0.261    11.3047700    0.0008  TRUE
## X-squared50   o3-Pro vs GPT-5-Medium                       0.138     3.7535757    0.0527  FALSE
## X-squared51   o3-Pro vs GPT-5-Low                          0.216     9.3018286    0.0023  TRUE
## X-squared52   o3-Pro vs GPT-5-Minimal                      0.406    31.7684094    0.0000  TRUE
## X-squared53   o3-Pro vs o4-mini-High                       0.247    18.7992418    0.0000  TRUE
## X-squared54   o3-Pro vs o4-mini-Medium                     0.305    28.0766993    0.0000  TRUE
## X-squared55   o3-Pro vs o3-GPT-Image-High                  0.211    19.3040615    0.0000  TRUE
## X-squared56   o3-Pro vs o3-GPT-Image-Medium                0.266    14.0713197    0.0002  TRUE
## X-squared57   GPT-5-Medium vs GPT-5-Low                    0.078     0.4606144    0.4973  FALSE
## X-squared58   GPT-5-Medium vs GPT-5-Minimal                0.268     7.5739980    0.0059  TRUE
## X-squared59   GPT-5-Medium vs o4-mini-High                 0.108     1.4939597    0.2216  FALSE
## X-squared60   GPT-5-Medium vs o4-mini-Medium               0.167     3.8069985    0.0510  FALSE
## X-squared61   GPT-5-Medium vs o3-GPT-Image-High            0.073     0.7723739    0.3795  FALSE
## X-squared62   GPT-5-Medium vs o3-GPT-Image-Medium          0.128     1.5123341    0.2188  FALSE
## X-squared63   GPT-5-Low vs GPT-5-Minimal                   0.190     3.6442900    0.0563  FALSE
## X-squared64   GPT-5-Low vs o4-mini-High                    0.031     0.0544585    0.8155  FALSE
## X-squared65   GPT-5-Low vs o4-mini-Medium                  0.089     0.9387856    0.3326  FALSE
## X-squared66   GPT-5-Low vs o3-GPT-Image-High              -0.004     0.0000000    1.0000  FALSE
## X-squared67   GPT-5-Low vs o3-GPT-Image-Medium             0.050     0.1365637    0.7117  FALSE
## X-squared68   GPT-5-Minimal vs o4-mini-High               -0.159     3.4681101    0.0626  FALSE
## X-squared69   GPT-5-Minimal vs o4-mini-Medium             -0.101     1.2846517    0.2570  FALSE
## X-squared70   GPT-5-Minimal vs o3-GPT-Image-High          -0.195     6.5374379    0.0106  TRUE
## X-squared71   GPT-5-Minimal vs o3-GPT-Image-Medium        -0.140     1.8579125    0.1729  FALSE
## X-squared72   o4-mini-High vs o4-mini-Medium               0.058     0.5967100    0.4398  FALSE
## X-squared73   o4-mini-High vs o3-GPT-Image-High           -0.035     0.2722063    0.6019  FALSE
## X-squared74   o4-mini-High vs o3-GPT-Image-Medium          0.019     0.0077047    0.9301  FALSE
## X-squared75   o4-mini-Medium vs o3-GPT-Image-High         -0.093     2.4427946    0.1181  FALSE
## X-squared76   o4-mini-Medium vs o3-GPT-Image-Medium       -0.039     0.1105138    0.7396  FALSE
## X-squared77   o3-GPT-Image-High vs o3-GPT-Image-Medium     0.055     0.3804382    0.5374  FALSE
```

Heatmap for Finke Reasoning Variations

```r
# Create matrix of p-values for Finke reasoning variations
finke_reasoning_models <- finke_reasoning_data$model
finke_reasoning_pval_matrix <- matrix(NA, nrow = length(finke_reasoning_models), ncol = length(finke_rea
rownames(finke_reasoning_pval_matrix) <- finke_reasoning_models
colnames(finke_reasoning_pval_matrix) <- finke_reasoning_models

for (i in 1:nrow(finke_reasoning_results)) {
  row_idx <- which(finke_reasoning_models == finke_reasoning_results$model1[i])
  col_idx <- which(finke_reasoning_models == finke_reasoning_results$model2[i])
  finke_reasoning_pval_matrix[row_idx, col_idx] <- finke_reasoning_results$p_value[i]
  finke_reasoning_pval_matrix[col_idx, row_idx] <- finke_reasoning_results$p_value[i]
}
# Set diagonal to NA
diag(finke_reasoning_pval_matrix) <- NA
# Set margins for better label display
par(mar = c(6, 6, 3, 2))
# Plot heatmap with same color palette
image(finke_reasoning_pval_matrix, axes = FALSE, col = col_palette,
      main = "P-values Heatmap - Finke Reasoning Variations")
axis(1, at = seq(0, 1, length.out = length(finke_reasoning_models)), labels = finke_reasoning_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_reasoning_models)), labels = finke_reasoning_models,
     las = 2, cex.axis = 0.8)
# Add gray color for diagonal
for (i in 1:length(finke_reasoning_models)) {
  x_pos <- (i - 1) / (length(finke_reasoning_models) - 1)
  y_pos <- (i - 1) / (length(finke_reasoning_models) - 1)
  rect(x_pos - 0.5 / (length(finke_reasoning_models) - 1), y_pos - 0.5 / (length(finke_reasoning_models
       x_pos + 0.5 / (length(finke_reasoning_models) - 1), y_pos + 0.5 / (length(finke_reasoning_models
       col = "gray80", border = NA)
}
# Add p-values to the plot
for (i in 1:nrow(finke_reasoning_pval_matrix)) {
  for (j in 1:ncol(finke_reasoning_pval_matrix)) {
    if (!is.na(finke_reasoning_pval_matrix[i, j])) {
      x_pos <- (j - 1) / (ncol(finke_reasoning_pval_matrix) - 1)
      y_pos <- (i - 1) / (nrow(finke_reasoning_pval_matrix) - 1)
      text(x_pos, y_pos, sprintf("%.3f", finke_reasoning_pval_matrix[i, j]), cex = 0.7)
    }
  }
}
```



Summary of Significant Differences - Finke Reasoning Variations

```r
# Count significant differences for Finke reasoning variations
finke_reasoning_sig_count <- sum(finke_reasoning_results$significant)
cat("Summary of Significant Differences - Finke Reasoning Variations:\n")
```

```
## Summary of Significant Differences - Finke Reasoning Variations:
```

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

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

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

```
##   Total comparisons: 78
```

```
cat("  Significant differences:", finke_reasoning_sig_count, "\n")
```

```
##   Significant differences: 29
```

```
cat("  Percentage significant:", round(finke_reasoning_sig_count / nrow(finke_reasoning_results) * 100,
```

```
##   Percentage significant: 37.2 %
```

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

```
## Significant Comparisons in Finke Reasoning Variations:
```

```
finke_reasoning_sig <- finke_reasoning_results[finke_reasoning_results$significant, c("comparison", "di
if (nrow(finke_reasoning_sig) > 0) {
  print(kable(finke_reasoning_sig, format = "simple", digits = 4))
} else {
  cat("  None\n")
}
```

```
##
##
##                 comparison                                 diff   p_value
## ------------  -----------------------------------  ---------  ---------
## X-squared3    Humans vs GPT-5-High                   -0.1360    0.0038
## X-squared4    Humans vs o3-Pro                       -0.1415    0.0002
## X-squared7    Humans vs GPT-5-Minimal                 0.2647    0.0001
## X-squared8    Humans vs o4-mini-High                  0.1052    0.0285
## X-squared9    Humans vs o4-mini-Medium                0.1634    0.0006
## X-squared10   Humans vs o3-GPT-Image-High             0.0699    0.0452
## X-squared14   o3-High vs GPT-5-High                  -0.1557    0.0071
## X-squared15   o3-High vs o3-Pro                      -0.1612    0.0014
## X-squared18   o3-High vs GPT-5-Minimal                0.2450    0.0016
## X-squared20   o3-High vs o4-mini-Medium               0.1437    0.0196
## X-squared24   o3-Medium vs GPT-5-High                -0.1926    0.0127
## X-squared25   o3-Medium vs o3-Pro                    -0.1981    0.0051
## X-squared28   o3-Medium vs GPT-5-Minimal              0.2080    0.0357
## X-squared34   o3-Low vs o3-Pro                       -0.1487    0.0367
## X-squared37   o3-Low vs GPT-5-Minimal                 0.2575    0.0083
## X-squared44   GPT-5-High vs GPT-5-Low                 0.2103    0.0065
## X-squared45   GPT-5-High vs GPT-5-Minimal             0.4007    0.0000
## X-squared46   GPT-5-High vs o4-mini-High              0.2412    0.0002
## X-squared47   GPT-5-High vs o4-mini-Medium            0.2994    0.0000
## X-squared48   GPT-5-High vs o3-GPT-Image-High         0.2059    0.0002
## X-squared49   GPT-5-High vs o3-GPT-Image-Medium       0.2606    0.0008
## X-squared51   o3-Pro vs GPT-5-Low                     0.2157    0.0023
## X-squared52   o3-Pro vs GPT-5-Minimal                 0.4061    0.0000
## X-squared53   o3-Pro vs o4-mini-High                  0.2466    0.0000
## X-squared54   o3-Pro vs o4-mini-Medium                0.3048    0.0000
## X-squared55   o3-Pro vs o3-GPT-Image-High             0.2114    0.0000
## X-squared56   o3-Pro vs o3-GPT-Image-Medium           0.2661    0.0002
## X-squared58   GPT-5-Medium vs GPT-5-Minimal           0.2679    0.0059
## X-squared70   GPT-5-Minimal vs o3-GPT-Image-High     -0.1948    0.0106
```

48 Novel

```
# Test all combinations for 48 Novel reasoning variations
novel_48_reasoning_results <- test_all_combinations(novel_reasoning_data, "48 Novel Reasoning Variations
# Display results
cat("All Pairwise Comparisons for 48 Novel Reasoning Variations:\n")
```

```
## All Pairwise Comparisons for 48 Novel Reasoning Variations:
```

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

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

```
for (i in 1:nrow(novel_48_reasoning_results)) {
  cat("\n", novel_48_reasoning_results$comparison[i], "\n")
  cat(paste(rep("-", 40), collapse = ""), "\n")
  cat("Proportions: ", round(novel_48_reasoning_results$prop1[i], 3), " vs ",
      round(novel_48_reasoning_results$prop2[i], 3), "\n")
  cat("Difference: ", round(novel_48_reasoning_results$diff[i], 3), "\n")
  cat("Chi-squared: ", round(novel_48_reasoning_results$chi_squared[i], 3), "\n")
  cat("Degrees of freedom: ", round(novel_48_reasoning_results$df[i], 3), "\n")
  cat("P-value: ", format(novel_48_reasoning_results$p_value[i], scientific = FALSE, digits = 4), "\n")
  cat("95% CI: [", round(novel_48_reasoning_results$ci_lower[i], 3), ", ",
      round(novel_48_reasoning_results$ci_upper[i], 3), "]\n")
  cat("Significant: ", ifelse(novel_48_reasoning_results$significant[i], "YES (p < 0.05)", "NO"), "\n")
}
```

```
##
##  Humans vs o3-High
## ----------------------------------------
## Proportions:  0.526  vs  0.649
## Difference:  -0.123
## Chi-squared:  38.861
## Degrees of freedom:  1
## P-value:  0.0000000004552
## 95% CI: [ -0.161 ,  -0.086 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o3-Medium
## ----------------------------------------
## Proportions:  0.526  vs  0.562
## Difference:  -0.036
## Chi-squared:  1.055
## Degrees of freedom:  1
## P-value:  0.3043
## 95% CI: [ -0.102 ,  0.03 ]
## Significant:  NO
##
##  Humans vs o3-Low
## ----------------------------------------
## Proportions:  0.526  vs  0.518
## Difference:  0.008
## Chi-squared:  0.027
## Degrees of freedom:  1
## P-value:  0.8702
## 95% CI: [ -0.059 ,  0.074 ]
## Significant:  NO
##
##  Humans vs GPT-5-High
## ----------------------------------------
## Proportions:  0.526  vs  0.646
## Difference:  -0.12
## Chi-squared:  25.084
## Degrees of freedom:  1
## P-value:  0.0000005489
## 95% CI: [ -0.165 ,  -0.074 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o3-Pro
## ----------------------------------------
## Proportions:  0.526  vs  0.585
## Difference:  -0.059
## Chi-squared:  2.995
## Degrees of freedom:  1
## P-value:  0.08352
## 95% CI: [ -0.125 ,  0.007 ]
## Significant:  NO
##
##  Humans vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.526  vs  0.64
## Difference:  -0.114
## Chi-squared:  33.112
## Degrees of freedom:  1
## P-value:  0.000000008702
## 95% CI: [ -0.152 ,  -0.076 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.526  vs  0.493
## Difference:  0.033
## Chi-squared:  0.881
## Degrees of freedom:  1
## P-value:  0.3479
## 95% CI: [ -0.034 ,  0.1 ]
## Significant:  NO
##
##  Humans vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.526  vs  0.418
## Difference:  0.108
## Chi-squared:  10.381
## Degrees of freedom:  1
## P-value:  0.001273
## 95% CI: [ 0.042 ,  0.174 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o4-mini-High
## ----------------------------------------
## Proportions:  0.526  vs  0.532
## Difference:  -0.006
## Chi-squared:  0.035
## Degrees of freedom:  1
## P-value:  0.8507
## 95% CI: [ -0.053 ,  0.042 ]
## Significant:  NO
##
##  Humans vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.526  vs  0.495
## Difference:  0.031
## Chi-squared:  1.563
## Degrees of freedom:  1
## P-value:  0.2112
## 95% CI: [ -0.017 ,  0.078 ]
## Significant:  NO
##
##  Humans vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.526  vs  0.552
## Difference:  -0.026
## Chi-squared:  2.115
## Degrees of freedom:  1
## P-value:  0.1458
## 95% CI: [ -0.06 ,  0.009 ]
## Significant:  NO
##
##  Humans vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.526  vs  0.559
## Difference:  -0.033
## Chi-squared:  0.897
## Degrees of freedom:  1
## P-value:  0.3435
## 95% CI: [ -0.1 ,  0.033 ]
## Significant:  NO
##
##  o3-High vs o3-Medium
## ----------------------------------------
## Proportions:  0.649  vs  0.562
## Difference:  0.087
## Chi-squared:  5.52
## Degrees of freedom:  1
## P-value:  0.01881
## 95% CI: [ 0.013 ,  0.162 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o3-Low
## ----------------------------------------
## Proportions:  0.649  vs  0.518
## Difference:  0.131
## Chi-squared:  12.5
## Degrees of freedom:  1
## P-value:  0.0004069
## 95% CI: [ 0.056 ,  0.206 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs GPT-5-High
## ----------------------------------------
## Proportions:  0.649  vs  0.646
## Difference:  0.004
## Chi-squared:  0.005
## Degrees of freedom:  1
## P-value:  0.9437
## 95% CI: [ -0.053 ,  0.061 ]
## Significant:  NO
##
##  o3-High vs o3-Pro
## ----------------------------------------
## Proportions:  0.649  vs  0.585
## Difference:  0.064
## Chi-squared:  2.936
## Degrees of freedom:  1
## P-value:  0.08663
## 95% CI: [ -0.01 ,  0.139 ]
## Significant:  NO
##
##  o3-High vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.649  vs  0.64
## Difference:  0.009
## Chi-squared:  0.101
## Degrees of freedom:  1
## P-value:  0.7504
## 95% CI: [ -0.041 ,  0.06 ]
## Significant:  NO
##
##  o3-High vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.649  vs  0.493
## Difference:  0.156
## Chi-squared:  17.859
## Degrees of freedom:  1
## P-value:  0.00002378
## 95% CI: [ 0.081 ,  0.231 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.649  vs  0.418
## Difference:  0.231
## Chi-squared:  38.969
## Degrees of freedom:  1
## P-value:  0.0000000004307
## 95% CI: [ 0.157 ,  0.306 ]
## Significant:  YES (p < 0.05)
##
```

```
##  o3-High vs o4-mini-High
## ----------------------------------------
## Proportions:  0.649  vs  0.532
## Difference:  0.118
## Chi-squared:  16.191
## Degrees of freedom:  1
## P-value:  0.00005726
## 95% CI: [ 0.059 ,  0.176 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.649  vs  0.495
## Difference:  0.154
## Chi-squared:  27.6
## Degrees of freedom:  1
## P-value:  0.0000001492
## 95% CI: [ 0.096 ,  0.213 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.649  vs  0.552
## Difference:  0.098
## Chi-squared:  15.818
## Degrees of freedom:  1
## P-value:  0.00006973
## 95% CI: [ 0.049 ,  0.146 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.649  vs  0.559
## Difference:  0.09
## Chi-squared:  5.863
## Degrees of freedom:  1
## P-value:  0.01546
## 95% CI: [ 0.015 ,  0.165 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs o3-Low
## ----------------------------------------
## Proportions:  0.562  vs  0.518
## Difference:  0.043
## Chi-squared:  0.746
## Degrees of freedom:  1
## P-value:  0.3877
## 95% CI: [ -0.05 ,  0.137 ]
## Significant:  NO
##
##  o3-Medium vs GPT-5-High
## ----------------------------------------
## Proportions:  0.562  vs  0.646
## Difference:  -0.084
## Chi-squared:  4.401
## Degrees of freedom:  1
## P-value:  0.03593
## 95% CI: [ -0.163 ,  -0.005 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs o3-Pro
## ----------------------------------------
## Proportions:  0.562  vs  0.585
## Difference:  -0.023
## Chi-squared:  0.176
## Degrees of freedom:  1
## P-value:  0.6747
## 95% CI: [ -0.116 ,  0.07 ]
## Significant:  NO
##
##  o3-Medium vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.562  vs  0.64
## Difference:  -0.078
## Chi-squared:  4.328
## Degrees of freedom:  1
## P-value:  0.03749
## 95% CI: [ -0.153 ,  -0.003 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.562  vs  0.493
## Difference:  0.069
## Chi-squared:  2.021
## Degrees of freedom:  1
## P-value:  0.1551
## 95% CI: [ -0.024 ,  0.162 ]
## Significant:  NO
##
##  o3-Medium vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.562  vs  0.418
## Difference:  0.144
## Chi-squared:  9.397
## Degrees of freedom:  1
## P-value:  0.002173
## 95% CI: [ 0.051 ,  0.237 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs o4-mini-High
## ----------------------------------------
## Proportions:  0.562  vs  0.532
## Difference:  0.03
## Chi-squared:  0.477
## Degrees of freedom:  1
## P-value:  0.4896
## 95% CI: [ -0.05 ,  0.11 ]
## Significant:  NO
##
##  o3-Medium vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.562  vs  0.495
## Difference:  0.067
## Chi-squared:  2.588
## Degrees of freedom:  1
## P-value:  0.1077
## 95% CI: [ -0.014 ,  0.147 ]
## Significant:  NO
##
##  o3-Medium vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.562  vs  0.552
## Difference:  0.01
## Chi-squared:  0.043
## Degrees of freedom:  1
## P-value:  0.8349
## 95% CI: [ -0.063 ,  0.083 ]
## Significant:  NO
##
##  o3-Medium vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.562  vs  0.559
## Difference:  0.003
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.089 ,  0.094 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-High
## ----------------------------------------
## Proportions:  0.518  vs  0.646
## Difference:  -0.127
## Chi-squared:  10.288
## Degrees of freedom:  1
## P-value:  0.001339
## 95% CI: [ -0.207 ,  -0.048 ]
## Significant:  YES (p < 0.05)
##
##  o3-Low vs o3-Pro
## ----------------------------------------
## Proportions:  0.518  vs  0.585
## Difference:  -0.067
## Chi-squared:  1.89
## Degrees of freedom:  1
## P-value:  0.1692
## 95% CI: [ -0.16 ,  0.026 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.518  vs  0.64
## Difference:  -0.121
## Chi-squared:  10.659
## Degrees of freedom:  1
## P-value:  0.001095
## 95% CI: [ -0.197 ,  -0.046 ]
## Significant:  YES (p < 0.05)
##
##  o3-Low vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.518  vs  0.493
## Difference:  0.025
## Chi-squared:  0.218
## Degrees of freedom:  1
## P-value:  0.6403
## 95% CI: [ -0.068 ,  0.119 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.518  vs  0.418
## Difference:  0.101
## Chi-squared:  4.481
## Degrees of freedom:  1
## P-value:  0.03427
## 95% CI: [ 0.008 ,  0.194 ]
## Significant:  YES (p < 0.05)
##
##  o3-Low vs o4-mini-High
## ----------------------------------------
## Proportions:  0.518  vs  0.532
## Difference:  -0.013
## Chi-squared:  0.064
## Degrees of freedom:  1
## P-value:  0.8
## 95% CI: [ -0.094 ,  0.067 ]
## Significant:  NO
##
##  o3-Low vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.518  vs  0.495
## Difference:  0.023
## Chi-squared:  0.258
## Degrees of freedom:  1
## P-value:  0.6113
## 95% CI: [ -0.057 ,  0.104 ]
## Significant:  NO
##
##  o3-Low vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.518  vs  0.552
## Difference:  -0.033
## Chi-squared:  0.734
## Degrees of freedom:  1
## P-value:  0.3917
## 95% CI: [ -0.107 ,  0.04 ]
## Significant:  NO
##
##  o3-Low vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.518  vs  0.559
## Difference:  -0.041
## Chi-squared:  0.649
## Degrees of freedom:  1
## P-value:  0.4203
## 95% CI: [ -0.134 ,  0.052 ]
## Significant:  NO
##
##  GPT-5-High vs o3-Pro
## ----------------------------------------
## Proportions:  0.646  vs  0.585
## Difference:  0.061
## Chi-squared:  2.256
## Degrees of freedom:  1
## P-value:  0.1331
## 95% CI: [ -0.018 ,  0.139 ]
## Significant:  NO
##
##  GPT-5-High vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.646  vs  0.64
## Difference:  0.006
## Chi-squared:  0.02
## Degrees of freedom:  1
## P-value:  0.8887
## 95% CI: [ -0.051 ,  0.063 ]
## Significant:  NO
##
##  GPT-5-High vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.646  vs  0.493
## Difference:  0.153
## Chi-squared:  14.846
## Degrees of freedom:  1
## P-value:  0.0001166
## 95% CI: [ 0.073 ,  0.232 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.646  vs  0.418
## Difference:  0.228
## Chi-squared:  32.938
## Degrees of freedom:  1
## P-value:  0.000000009513
## 95% CI: [ 0.149 ,  0.307 ]
## Significant:  YES (p < 0.05)
##
```

```
##  GPT-5-High vs o4-mini-High
## ----------------------------------------
## Proportions:  0.646  vs  0.532
## Difference:  0.114
## Chi-squared:  12.427
## Degrees of freedom:  1
## P-value:  0.0004231
## 95% CI: [ 0.05 ,  0.178 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.646  vs  0.495
## Difference:  0.15
## Chi-squared:  21.544
## Degrees of freedom:  1
## P-value:  0.000003457
## 95% CI: [ 0.086 ,  0.214 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.646  vs  0.552
## Difference:  0.094
## Chi-squared:  11.198
## Degrees of freedom:  1
## P-value:  0.0008187
## 95% CI: [ 0.039 ,  0.148 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.646  vs  0.559
## Difference:  0.086
## Chi-squared:  4.688
## Degrees of freedom:  1
## P-value:  0.03037
## 95% CI: [ 0.007 ,  0.165 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.585  vs  0.64
## Difference:  -0.055
## Chi-squared:  2.09
## Degrees of freedom:  1
## P-value:  0.1483
## 95% CI: [ -0.129 ,  0.019 ]
## Significant:  NO
##
##  o3-Pro vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.585  vs  0.493
## Difference:  0.092
## Chi-squared:  3.732
## Degrees of freedom:  1
## P-value:  0.05338
## 95% CI: [ -0.001 ,  0.185 ]
## Significant:  NO
##
##  o3-Pro vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.585  vs  0.418
## Difference:  0.167
## Chi-squared:  12.754
## Degrees of freedom:  1
## P-value:  0.0003552
## 95% CI: [ 0.075 ,  0.26 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o4-mini-High
## ----------------------------------------
## Proportions:  0.585  vs  0.532
## Difference:  0.053
## Chi-squared:  1.637
## Degrees of freedom:  1
## P-value:  0.2007
## 95% CI: [ -0.026 ,  0.133 ]
## Significant:  NO
##
##  o3-Pro vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.585  vs  0.495
## Difference:  0.09
## Chi-squared:  4.82
## Degrees of freedom:  1
## P-value:  0.02813
## 95% CI: [ 0.01 ,  0.17 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.585  vs  0.552
## Difference:  0.033
## Chi-squared:  0.729
## Degrees of freedom:  1
## P-value:  0.3933
## 95% CI: [ -0.039 ,  0.106 ]
## Significant:  NO
##
##  o3-Pro vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.585  vs  0.559
## Difference:  0.026
## Chi-squared:  0.228
## Degrees of freedom:  1
## P-value:  0.6328
## 95% CI: [ -0.067 ,  0.118 ]
## Significant:  NO
##
##  GPT-5-Medium vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.64  vs  0.493
## Difference:  0.147
## Chi-squared:  15.639
## Degrees of freedom:  1
## P-value:  0.00007667
## 95% CI: [ 0.072 ,  0.222 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.64  vs  0.418
## Difference:  0.222
## Chi-squared:  35.644
## Degrees of freedom:  1
## P-value:  0.000000002369
## 95% CI: [ 0.148 ,  0.296 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs o4-mini-High
## ----------------------------------------
## Proportions:  0.64  vs  0.532
## Difference:  0.108
## Chi-squared:  13.607
## Degrees of freedom:  1
## P-value:  0.0002253
## 95% CI: [ 0.05 ,  0.167 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.64  vs  0.495
## Difference:  0.145
## Chi-squared:  24.2
## Degrees of freedom:  1
## P-value:  0.0000008685
## 95% CI: [ 0.086 ,  0.203 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.64  vs  0.552
## Difference:  0.088
## Chi-squared:  12.838
## Degrees of freedom:  1
## P-value:  0.0003397
## 95% CI: [ 0.04 ,  0.136 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.64  vs  0.559
## Difference:  0.081
## Chi-squared:  4.633
## Degrees of freedom:  1
## P-value:  0.03137
## 95% CI: [ 0.006 ,  0.155 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Low vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.493  vs  0.418
## Difference:  0.075
## Chi-squared:  2.435
## Degrees of freedom:  1
## P-value:  0.1187
## 95% CI: [ -0.018 ,  0.168 ]
## Significant:  NO
##
##  GPT-5-Low vs o4-mini-High
## ----------------------------------------
## Proportions:  0.493  vs  0.532
## Difference:  -0.039
## Chi-squared:  0.807
## Degrees of freedom:  1
## P-value:  0.369
## 95% CI: [ -0.119 ,  0.042 ]
## Significant:  NO
##
##  GPT-5-Low vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.493  vs  0.495
## Difference:  -0.002
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.082 ,  0.077 ]
## Significant:  NO
##
##  GPT-5-Low vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.493  vs  0.552
## Difference:  -0.059
## Chi-squared:  2.448
## Degrees of freedom:  1
## P-value:  0.1177
## 95% CI: [ -0.132 ,  0.014 ]
## Significant:  NO
##
##  GPT-5-Low vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.493  vs  0.559
## Difference:  -0.066
## Chi-squared:  1.86
## Degrees of freedom:  1
## P-value:  0.1726
## 95% CI: [ -0.16 ,  0.027 ]
## Significant:  NO
##
##  GPT-5-Minimal vs o4-mini-High
## ----------------------------------------
## Proportions:  0.418  vs  0.532
## Difference:  -0.114
## Chi-squared:  7.828
## Degrees of freedom:  1
## P-value:  0.005144
## 95% CI: [ -0.194 ,  -0.034 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Minimal vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.418  vs  0.495
## Difference:  -0.077
## Chi-squared:  3.542
## Degrees of freedom:  1
## P-value:  0.05984
## 95% CI: [ -0.157 ,  0.003 ]
## Significant:  NO
##
##  GPT-5-Minimal vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.418  vs  0.552
## Difference:  -0.134
## Chi-squared:  13.288
## Degrees of freedom:  1
## P-value:  0.0002671
## 95% CI: [ -0.206 ,  -0.061 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Minimal vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.418  vs  0.559
## Difference:  -0.141
## Chi-squared:  9.048
## Degrees of freedom:  1
## P-value:  0.002629
## 95% CI: [ -0.234 ,  -0.049 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-High vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.532  vs  0.495
## Difference:  0.036
## Chi-squared:  1.127
## Degrees of freedom:  1
## P-value:  0.2884
## 95% CI: [ -0.029 ,  0.102 ]
## Significant:  NO
##
```

```
##  o4-mini-High vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.532  vs  0.552
## Difference:  -0.02
## Chi-squared:  0.451
## Degrees of freedom:  1
## P-value:  0.5017
## 95% CI: [ -0.076 ,  0.036 ]
## Significant:  NO
##
##  o4-mini-High vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.532  vs  0.559
## Difference:  -0.028
## Chi-squared:  0.39
## Degrees of freedom:  1
## P-value:  0.5325
## 95% CI: [ -0.108 ,  0.052 ]
## Significant:  NO
##
##  o4-mini-Medium vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.495  vs  0.552
## Difference:  -0.057
## Chi-squared:  3.894
## Degrees of freedom:  1
## P-value:  0.04845
## 95% CI: [ -0.113 ,  0 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-Medium vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.495  vs  0.559
## Difference:  -0.064
## Chi-squared:  2.378
## Degrees of freedom:  1
## P-value:  0.123
## 95% CI: [ -0.144 ,  0.016 ]
## Significant:  NO
##
##  o3-GPT-Image-High vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.552  vs  0.559
## Difference:  -0.007
## Chi-squared:  0.018
## Degrees of freedom:  1
## P-value:  0.8925
## 95% CI: [ -0.08 ,  0.065 ]
## Significant:  NO
```

```
# Summary table
novel_48_reasoning_summary <- novel_48_reasoning_results %>%
  select(comparison, diff, chi_squared, p_value, significant) %>%
  mutate(diff = round(diff, 3),
```

```r
         p_value = round(p_value, 4))
cat("\n\nSummary Table - 48 Novel Reasoning Variations:\n")
```

```
##
##
## Summary Table - 48 Novel Reasoning Variations:
```

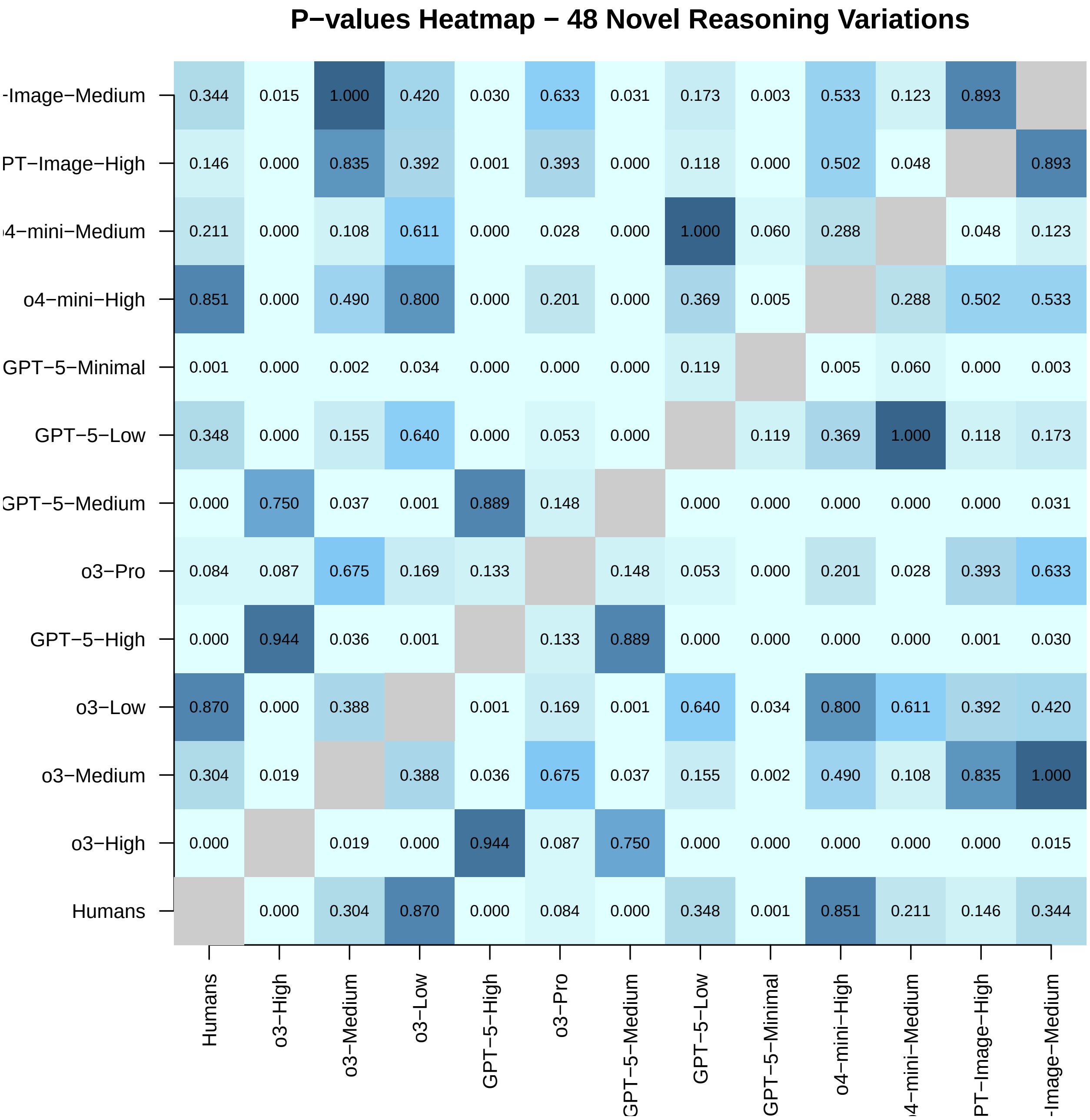
```r
print(kable(novel_48_reasoning_summary, format = "simple"))
```

```
##
##
##                comparison                                    diff   chi_squared   p_value  significan
## ------------  -----------------------------------------  -------  ------------  --------  ----------
## X-squared     Humans vs o3-High                           -0.123    38.8606815    0.0000  TRUE
## X-squared1    Humans vs o3-Medium                         -0.036     1.0550879    0.3043  FALSE
## X-squared2    Humans vs o3-Low                             0.008     0.0266943    0.8702  FALSE
## X-squared3    Humans vs GPT-5-High                        -0.120    25.0838266    0.0000  TRUE
## X-squared4    Humans vs o3-Pro                            -0.059     2.9949910    0.0835  FALSE
## X-squared5    Humans vs GPT-5-Medium                      -0.114    33.1116612    0.0000  TRUE
## X-squared6    Humans vs GPT-5-Low                          0.033     0.8811517    0.3479  FALSE
## X-squared7    Humans vs GPT-5-Minimal                      0.108    10.3814323    0.0013  TRUE
## X-squared8    Humans vs o4-mini-High                      -0.006     0.0354255    0.8507  FALSE
## X-squared9    Humans vs o4-mini-Medium                     0.031     1.5632859    0.2112  FALSE
## X-squared10   Humans vs o3-GPT-Image-High                 -0.026     2.1151564    0.1458  FALSE
## X-squared11   Humans vs o3-GPT-Image-Medium               -0.033     0.8971500    0.3435  FALSE
## X-squared12   o3-High vs o3-Medium                         0.087     5.5195207    0.0188  TRUE
## X-squared13   o3-High vs o3-Low                            0.131    12.5001282    0.0004  TRUE
## X-squared14   o3-High vs GPT-5-High                        0.004     0.0049901    0.9437  FALSE
## X-squared15   o3-High vs o3-Pro                            0.064     2.9359820    0.0866  FALSE
## X-squared16   o3-High vs GPT-5-Medium                      0.009     0.1011775    0.7504  FALSE
## X-squared17   o3-High vs GPT-5-Low                         0.156    17.8593660    0.0000  TRUE
## X-squared18   o3-High vs GPT-5-Minimal                     0.231    38.9685289    0.0000  TRUE
## X-squared19   o3-High vs o4-mini-High                      0.118    16.1910358    0.0001  TRUE
## X-squared20   o3-High vs o4-mini-Medium                    0.154    27.5997132    0.0000  TRUE
## X-squared21   o3-High vs o3-GPT-Image-High                 0.098    15.8181217    0.0001  TRUE
## X-squared22   o3-High vs o3-GPT-Image-Medium               0.090     5.8634324    0.0155  TRUE
## X-squared23   o3-Medium vs o3-Low                          0.043     0.7461811    0.3877  FALSE
## X-squared24   o3-Medium vs GPT-5-High                     -0.084     4.4006182    0.0359  TRUE
## X-squared25   o3-Medium vs o3-Pro                         -0.023     0.1761369    0.6747  FALSE
## X-squared26   o3-Medium vs GPT-5-Medium                   -0.078     4.3280310    0.0375  TRUE
## X-squared27   o3-Medium vs GPT-5-Low                       0.069     2.0210774    0.1551  FALSE
## X-squared28   o3-Medium vs GPT-5-Minimal                   0.144     9.3972334    0.0022  TRUE
## X-squared29   o3-Medium vs o4-mini-High                    0.030     0.4774938    0.4896  FALSE
## X-squared30   o3-Medium vs o4-mini-Medium                  0.067     2.5883714    0.1077  FALSE
## X-squared31   o3-Medium vs o3-GPT-Image-High               0.010     0.0434316    0.8349  FALSE
## X-squared32   o3-Medium vs o3-GPT-Image-Medium             0.003     0.0000000    1.0000  FALSE
## X-squared33   o3-Low vs GPT-5-High                        -0.127    10.2882273    0.0013  TRUE
## X-squared34   o3-Low vs o3-Pro                            -0.067     1.8901321    0.1692  FALSE
## X-squared35   o3-Low vs GPT-5-Medium                      -0.121    10.6590196    0.0011  TRUE
## X-squared36   o3-Low vs GPT-5-Low                          0.025     0.2182983    0.6403  FALSE
## X-squared37   o3-Low vs GPT-5-Minimal                      0.101     4.4810600    0.0343  TRUE
## X-squared38   o3-Low vs o4-mini-High                      -0.013     0.0641572    0.8000  FALSE
## X-squared39   o3-Low vs o4-mini-Medium                     0.023     0.2582995    0.6113  FALSE
## X-squared40   o3-Low vs o3-GPT-Image-High                 -0.033     0.7336947    0.3917  FALSE
## X-squared41   o3-Low vs o3-GPT-Image-Medium               -0.041     0.6494190    0.4203  FALSE
## X-squared42   GPT-5-High vs o3-Pro                         0.061     2.2561993    0.1331  FALSE
## X-squared43   GPT-5-High vs GPT-5-Medium                   0.006     0.0195799    0.8887  FALSE
## X-squared44   GPT-5-High vs GPT-5-Low                      0.153    14.8462373    0.0001  TRUE
## X-squared45   GPT-5-High vs GPT-5-Minimal                  0.228    32.9383855    0.0000  TRUE
## X-squared46   GPT-5-High vs o4-mini-High                   0.114    12.4274222    0.0004  TRUE
## X-squared47   GPT-5-High vs o4-mini-Medium                 0.150    21.5444489    0.0000  TRUE
## X-squared48   GPT-5-High vs o3-GPT-Image-High              0.094    11.1983901    0.0008  TRUE
## X-squared49   GPT-5-High vs o3-GPT-Image-Medium            0.086     4.6883219    0.0304  TRUE
## X-squared50   o3-Pro vs GPT-5-Medium                      -0.055     2.0901458    0.1483  FALSE
## X-squared51   o3-Pro vs GPT-5-Low                          0.092     3.7318677    0.0534  FALSE
## X-squared52   o3-Pro vs GPT-5-Minimal                      0.167    12.7541485    0.0004  TRUE
## X-squared53   o3-Pro vs o4-mini-High                       0.053     1.6374358    0.2007  FALSE
## X-squared54   o3-Pro vs o4-mini-Medium                     0.090     4.8198083    0.0281  TRUE
## X-squared55   o3-Pro vs o3-GPT-Image-High                  0.033     0.7286695    0.3933  FALSE
## X-squared56   o3-Pro vs o3-GPT-Image-Medium                0.026     0.2282248    0.6328  FALSE
## X-squared57   GPT-5-Medium vs GPT-5-Low                    0.147    15.6387810    0.0001  TRUE
## X-squared58   GPT-5-Medium vs GPT-5-Minimal                0.222    35.6439923    0.0000  TRUE
## X-squared59   GPT-5-Medium vs o4-mini-High                 0.108    13.6072593    0.0002  TRUE
## X-squared60   GPT-5-Medium vs o4-mini-Medium               0.145    24.1996546    0.0000  TRUE
## X-squared61   GPT-5-Medium vs o3-GPT-Image-High            0.088    12.8378194    0.0003  TRUE
## X-squared62   GPT-5-Medium vs o3-GPT-Image-Medium          0.081     4.6326164    0.0314  TRUE
## X-squared63   GPT-5-Low vs GPT-5-Minimal                   0.075     2.4345048    0.1187  FALSE
## X-squared64   GPT-5-Low vs o4-mini-High                   -0.039     0.8071628    0.3690  FALSE
## X-squared65   GPT-5-Low vs o4-mini-Medium                 -0.002     0.0000000    1.0000  FALSE
## X-squared66   GPT-5-Low vs o3-GPT-Image-High              -0.059     2.4476967    0.1177  FALSE
## X-squared67   GPT-5-Low vs o3-GPT-Image-Medium            -0.066     1.8598675    0.1726  FALSE
## X-squared68   GPT-5-Minimal vs o4-mini-High               -0.114     7.8282651    0.0051  TRUE
## X-squared69   GPT-5-Minimal vs o4-mini-Medium             -0.077     3.5417399    0.0598  FALSE
## X-squared70   GPT-5-Minimal vs o3-GPT-Image-High          -0.134    13.2879128    0.0003  TRUE
## X-squared71   GPT-5-Minimal vs o3-GPT-Image-Medium        -0.141     9.0483795    0.0026  TRUE
## X-squared72   o4-mini-High vs o4-mini-Medium               0.036     1.1271703    0.2884  FALSE
## X-squared73   o4-mini-High vs o3-GPT-Image-High           -0.020     0.4513398    0.5017  FALSE
## X-squared74   o4-mini-High vs o3-GPT-Image-Medium         -0.028     0.3895990    0.5325  FALSE
## X-squared75   o4-mini-Medium vs o3-GPT-Image-High         -0.057     3.8942781    0.0485  TRUE
## X-squared76   o4-mini-Medium vs o3-GPT-Image-Medium       -0.064     2.3783849    0.1230  FALSE
## X-squared77   o3-GPT-Image-High vs o3-GPT-Image-Medium    -0.007     0.0182536    0.8925  FALSE
```

Heatmap for 48 Novel Reasoning Variations

```
# Create matrix of p-values for 48 Novel reasoning variations
novel_48_reasoning_models <- novel_reasoning_data$model
novel_48_reasoning_pval_matrix <- matrix(NA, nrow = length(novel_48_reasoning_models), ncol = length(no
rownames(novel_48_reasoning_pval_matrix) <- novel_48_reasoning_models
colnames(novel_48_reasoning_pval_matrix) <- novel_48_reasoning_models

for (i in 1:nrow(novel_48_reasoning_results)) {
  row_idx <- which(novel_48_reasoning_models == novel_48_reasoning_results$model1[i])
  col_idx <- which(novel_48_reasoning_models == novel_48_reasoning_results$model2[i])
  novel_48_reasoning_pval_matrix[row_idx, col_idx] <- novel_48_reasoning_results$p_value[i]
  novel_48_reasoning_pval_matrix[col_idx, row_idx] <- novel_48_reasoning_results$p_value[i]
}
# Set diagonal to NA
diag(novel_48_reasoning_pval_matrix) <- NA
```

```
# Set margins for better label display
par(mar = c(6, 6, 3, 2))
# Plot heatmap with same color palette
image(novel_48_reasoning_pval_matrix, axes = FALSE, col = col_palette,
      main = "P-values Heatmap - 48 Novel Reasoning Variations")
axis(1, at = seq(0, 1, length.out = length(novel_48_reasoning_models)), labels = novel_48_reasoning_mode
     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_48_reasoning_models)), labels = novel_48_reasoning_mode
     las = 2, cex.axis = 0.8)
# Add gray color for diagonal
for (i in 1:length(novel_48_reasoning_models)) {
  x_pos <- (i - 1) / (length(novel_48_reasoning_models) - 1)
  y_pos <- (i - 1) / (length(novel_48_reasoning_models) - 1)
  rect(x_pos - 0.5 / (length(novel_48_reasoning_models) - 1), y_pos - 0.5 / (length(novel_48_reasoning_
       x_pos + 0.5 / (length(novel_48_reasoning_models) - 1), y_pos + 0.5 / (length(novel_48_reasoning_
       col = "gray80", border = NA)
}
# Add p-values to the plot
for (i in 1:nrow(novel_48_reasoning_pval_matrix)) {
  for (j in 1:ncol(novel_48_reasoning_pval_matrix)) {
    if (!is.na(novel_48_reasoning_pval_matrix[i, j])) {
      x_pos <- (j - 1) / (ncol(novel_48_reasoning_pval_matrix) - 1)
      y_pos <- (i - 1) / (nrow(novel_48_reasoning_pval_matrix) - 1)
      text(x_pos, y_pos, sprintf("%.3f", novel_48_reasoning_pval_matrix[i, j]), cex = 0.7)
    }
  }
}
```



P−values Heatmap − 48 Novel Reasoning Variations

	Humans	o3−High	o3−Medium	o3−Low	GPT−5−High	o3−Pro	GPT−5−Medium	GPT−5−Low	GPT−5−Minimal	o4−mini−High	4−mini−Medium	PT−Image−High	-Image−Medium
-Image−Medium	0.344	0.015	1.000	0.420	0.030	0.633	0.031	0.173	0.003	0.533	0.123	0.893	
PT−Image−High	0.146	0.000	0.835	0.392	0.001	0.393	0.000	0.118	0.000	0.502	0.048		0.893
4−mini−Medium	0.211	0.000	0.108	0.611	0.000	0.028	0.000	1.000	0.060	0.288		0.048	0.123
o4−mini−High	0.851	0.000	0.490	0.800	0.000	0.201	0.000	0.369	0.005		0.288	0.502	0.533
GPT−5−Minimal	0.001	0.000	0.002	0.034	0.000	0.000	0.000	0.119		0.005	0.060	0.000	0.003
GPT−5−Low	0.348	0.000	0.155	0.640	0.000	0.053	0.000		0.119	0.369	1.000	0.118	0.173
GPT−5−Medium	0.000	0.750	0.037	0.001	0.889	0.148		0.000	0.000	0.000	0.000	0.000	0.031
o3−Pro	0.084	0.087	0.675	0.169	0.133		0.148	0.053	0.000	0.201	0.028	0.393	0.633
GPT−5−High	0.000	0.944	0.036	0.001		0.133	0.889	0.000	0.000	0.000	0.000	0.001	0.030
o3−Low	0.870	0.000	0.388		0.001	0.169	0.001	0.640	0.034	0.800	0.611	0.392	0.420
o3−Medium	0.304	0.019		0.388	0.036	0.675	0.037	0.155	0.002	0.490	0.108	0.835	1.000
o3−High	0.000		0.019	0.000	0.944	0.087	0.750	0.000	0.000	0.000	0.000	0.000	0.015
Humans		0.000	0.304	0.870	0.000	0.084	0.000	0.348	0.001	0.851	0.211	0.146	0.344

Summary of Significant Differences - 48 Novel Reasoning Variations

```r
# Count significant differences for 48 Novel reasoning variations
novel_48_reasoning_sig_count <- sum(novel_48_reasoning_results$significant)
cat("Summary of Significant Differences - 48 Novel Reasoning Variations:\n")
```

```
## Summary of Significant Differences - 48 Novel Reasoning Variations:
```

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

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

```r
cat("  Total comparisons:", nrow(novel_48_reasoning_results), "\n")
```

```
##   Total comparisons: 78
```

```
cat("  Significant differences:", novel_48_reasoning_sig_count, "\n")
```

```
##   Significant differences: 36
```

```
cat("  Percentage significant:", round(novel_48_reasoning_sig_count / nrow(novel_48_reasoning_results)
```

```
##   Percentage significant: 46.2 %
```

```
# Show which comparisons are significant
cat("Significant Comparisons in 48 Novel Reasoning Variations:\n")
```

```
## Significant Comparisons in 48 Novel Reasoning Variations:
```

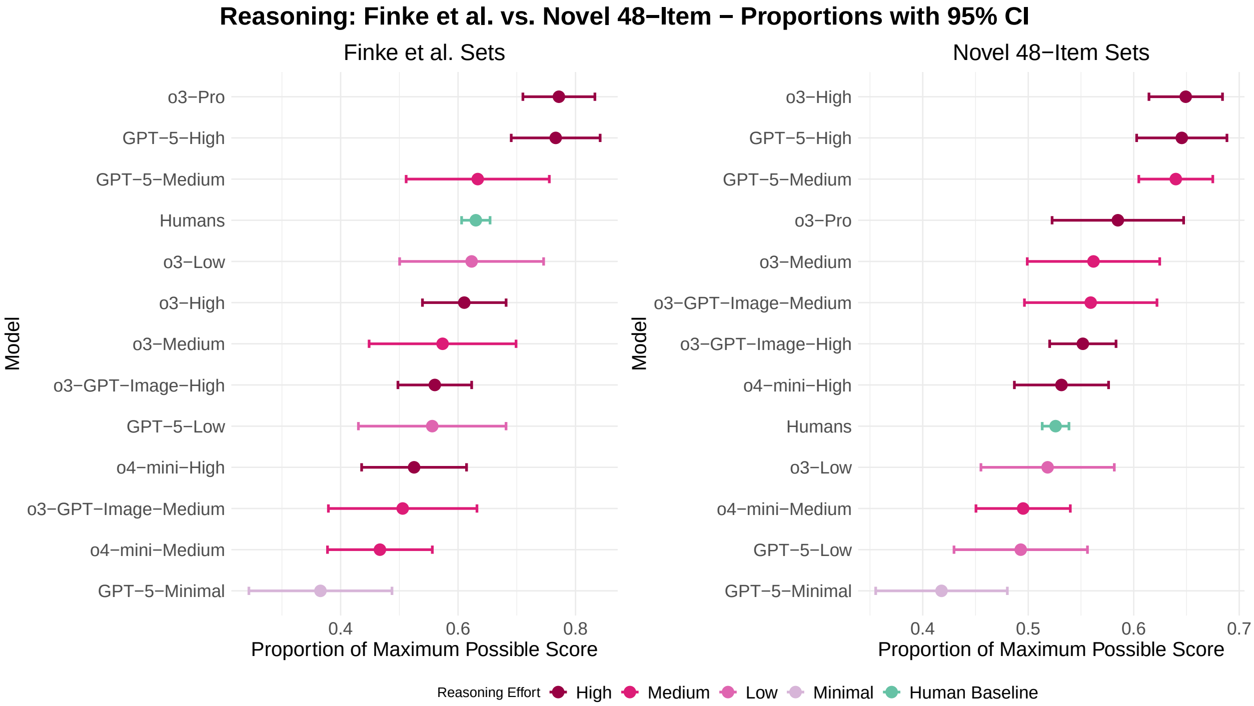
```
novel_48_reasoning_sig <- novel_48_reasoning_results[novel_48_reasoning_results$significant, c("compari
if (nrow(novel_48_reasoning_sig) > 0) {
  print(kable(novel_48_reasoning_sig, format = "simple", digits = 4))
} else {
  cat("  None\n")
}
```

```
##
##
##                comparison                                diff   p_value
## ------------  -------------------------------------  ---------  --------
## X-squared     Humans vs o3-High                        -0.1234    0.0000
## X-squared3    Humans vs GPT-5-High                     -0.1196    0.0000
## X-squared5    Humans vs GPT-5-Medium                   -0.1140    0.0000
## X-squared7    Humans vs GPT-5-Minimal                   0.1081    0.0013
## X-squared12   o3-High vs o3-Medium                      0.0874    0.0188
## X-squared13   o3-High vs o3-Low                         0.1309    0.0004
## X-squared17   o3-High vs GPT-5-Low                      0.1564    0.0000
## X-squared18   o3-High vs GPT-5-Minimal                  0.2315    0.0000
## X-squared19   o3-High vs o4-mini-High                   0.1178    0.0001
## X-squared20   o3-High vs o4-mini-Medium                 0.1541    0.0000
## X-squared21   o3-High vs o3-GPT-Image-High              0.0975    0.0001
## X-squared22   o3-High vs o3-GPT-Image-Medium            0.0901    0.0155
## X-squared24   o3-Medium vs GPT-5-High                  -0.0837    0.0359
## X-squared26   o3-Medium vs GPT-5-Medium                -0.0780    0.0375
## X-squared28   o3-Medium vs GPT-5-Minimal                0.1441    0.0022
## X-squared33   o3-Low vs GPT-5-High                     -0.1272    0.0013
## X-squared35   o3-Low vs GPT-5-Medium                   -0.1215    0.0011
## X-squared37   o3-Low vs GPT-5-Minimal                   0.1006    0.0343
## X-squared44   GPT-5-High vs GPT-5-Low                   0.1527    0.0001
## X-squared45   GPT-5-High vs GPT-5-Minimal               0.2278    0.0000
## X-squared46   GPT-5-High vs o4-mini-High                0.1141    0.0004
## X-squared47   GPT-5-High vs o4-mini-Medium              0.1504    0.0000
## X-squared48   GPT-5-High vs o3-GPT-Image-High           0.0938    0.0008
## X-squared49   GPT-5-High vs o3-GPT-Image-Medium         0.0863    0.0304
## X-squared52   o3-Pro vs GPT-5-Minimal                   0.1672    0.0004
## X-squared54   o3-Pro vs o4-mini-Medium                  0.0898    0.0281
## X-squared57   GPT-5-Medium vs GPT-5-Low                 0.1470    0.0001
## X-squared58   GPT-5-Medium vs GPT-5-Minimal             0.2221    0.0000
## X-squared59   GPT-5-Medium vs o4-mini-High              0.1084    0.0002
## X-squared60   GPT-5-Medium vs o4-mini-Medium            0.1447    0.0000
## X-squared61   GPT-5-Medium vs o3-GPT-Image-High         0.0881    0.0003
## X-squared62   GPT-5-Medium vs o3-GPT-Image-Medium      0.0807    0.0314
## X-squared68   GPT-5-Minimal vs o4-mini-High           -0.1137    0.0051
## X-squared70   GPT-5-Minimal vs o3-GPT-Image-High      -0.1340    0.0003
## X-squared71   GPT-5-Minimal vs o3-GPT-Image-Medium    -0.1414    0.0026
## X-squared75   o4-mini-Medium vs o3-GPT-Image-High     -0.0566    0.0485
```

Visualization of Finke and Novel Reasoning Variations

```
# Plot proportions with confidence intervals for Finke reasoning variations
finke_reasoning_plot <- ggplot(finke_reasoning_data, aes(x = reorder(model, proportion), y = proportion
  geom_point(size = 4, aes(color = color)) +
  geom_errorbar(aes(ymin = proportion - 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    ymax = proportion + 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    color = color),
                width = 0.2, size = 1) +
  coord_flip() +
  theme_minimal() +
  labs(subtitle = "Finke et al. Sets",
       x = "Model",
       y = "Proportion of Maximum Possible Score") +
  theme(plot.subtitle = element_text(hjust = 0.5, size = 18),
        axis.text = element_text(size = 14),
        axis.title = element_text(size = 16),
        legend.text = element_text(size = 14)) +
  scale_color_manual(
    values = c("#980043", "#dd1c77", "#df65b0", "#d7b5d8", "#66c2a5"),
    name = "Reasoning Effort",
    breaks = c("#980043", "#dd1c77", "#df65b0", "#d7b5d8", "#66c2a5"),
    labels = c("High", "Medium", "Low", "Minimal", "Human Baseline")
  )

# Plot proportions with confidence intervals for 48 Novel reasoning variations
novel_48_reasoning_plot <- ggplot(novel_reasoning_data, aes(x = reorder(model, proportion), y = proporti
  geom_point(size = 4, aes(color = color)) +
  geom_errorbar(aes(ymin = proportion - 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    ymax = proportion + 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    color = color),
                width = 0.2, size = 1) +
  coord_flip() +
  theme_minimal() +
  labs(subtitle = "Novel 48-Item Sets",
       x = "Model",
       y = "Proportion of Maximum Possible Score") +
  theme(plot.subtitle = element_text(hjust = 0.5, size = 18),
        axis.text = element_text(size = 14),
        axis.title = element_text(size = 16),
        legend.text = element_text(size = 14)) +
  scale_color_manual(
    values = c("#980043", "#dd1c77", "#df65b0", "#d7b5d8", "#66c2a5"),
    name = "Reasoning Effort",
    breaks = c("#980043", "#dd1c77", "#df65b0", "#d7b5d8", "#66c2a5"),
    labels = c("High", "Medium", "Low", "Minimal", "Human Baseline")
  )
```

```r
combined_reasoning_plot <- ((finke_reasoning_plot + novel_48_reasoning_plot) +
  plot_layout(ncol = 2, guides = "collect") +
  plot_annotation(title = "Reasoning: Finke et al. vs. Novel 48-Item - Proportions with 95% CI")) &
  theme(plot.title = element_text(hjust = 0.5, size = 20, face = "bold"), legend.position = "bottom")
print(combined_reasoning_plot)
```



Combined Summary of Reasoning Variations

```r
combined_reasoning_results <- test_all_combinations(collapsed_reasoning_data, "Combined Reasoning Varia
# Display results
cat("All Pairwise Comparisons for Combined Reasoning Variations:\n")
```

```
## All Pairwise Comparisons for Combined Reasoning Variations:
```

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

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

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

```
##
##  Humans vs o3-High
## ----------------------------------------
## Proportions:  0.547  vs  0.642
## Difference:  -0.094
## Chi-squared:  28.631
## Degrees of freedom:  1
## P-value:  0.00000008757
## 95% CI: [ -0.128 ,  -0.06 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o3-Medium
## ----------------------------------------
## Proportions:  0.547  vs  0.564
## Difference:  -0.017
## Chi-squared:  0.273
## Degrees of freedom:  1
## P-value:  0.6014
## 95% CI: [ -0.076 ,  0.042 ]
## Significant:  NO
##
##  Humans vs o3-Low
## ----------------------------------------
## Proportions:  0.547  vs  0.539
## Difference:  0.008
## Chi-squared:  0.043
## Degrees of freedom:  1
## P-value:  0.8349
## 95% CI: [ -0.051 ,  0.067 ]
## Significant:  NO
##
##  Humans vs GPT-5-High
## ----------------------------------------
## Proportions:  0.547  vs  0.67
## Difference:  -0.123
## Chi-squared:  33.302
## Degrees of freedom:  1
## P-value:  0.00000000789
## 95% CI: [ -0.163 ,  -0.082 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o3-Pro
## ----------------------------------------
## Proportions:  0.547  vs  0.666
## Difference:  -0.119
## Chi-squared:  45.76
## Degrees of freedom:  1
## P-value:  0.00000000001336
## 95% CI: [ -0.153 ,  -0.086 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.547  vs  0.595
## Difference:  -0.048
## Chi-squared:  2.441
## Degrees of freedom:  1
## P-value:  0.1182
## 95% CI: [ -0.106 ,  0.011 ]
## Significant:  NO
##
##  Humans vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.547  vs  0.506
## Difference:  0.042
## Chi-squared:  1.854
## Degrees of freedom:  1
## P-value:  0.1733
## 95% CI: [ -0.018 ,  0.101 ]
## Significant:  NO
##
##  Humans vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.547  vs  0.407
## Difference:  0.14
## Chi-squared:  22.151
## Degrees of freedom:  1
## P-value:  0.00000252
## 95% CI: [ 0.081 ,  0.198 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o4-mini-High
## ----------------------------------------
## Proportions:  0.547  vs  0.53
## Difference:  0.017
## Chi-squared:  0.577
## Degrees of freedom:  1
## P-value:  0.4477
## 95% CI: [ -0.025 ,  0.059 ]
## Significant:  NO
##
##  Humans vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.547  vs  0.49
## Difference:  0.058
## Chi-squared:  7.209
## Degrees of freedom:  1
## P-value:  0.007255
## 95% CI: [ 0.015 ,  0.1 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.547  vs  0.553
## Difference:  -0.006
## Chi-squared:  0.143
## Degrees of freedom:  1
## P-value:  0.7057
## 95% CI: [ -0.037 ,  0.024 ]
## Significant:  NO
##
##  Humans vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.547  vs  0.549
## Difference:  -0.001
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.06 ,  0.057 ]
## Significant:  NO
##
##  o3-High vs o3-Medium
## ----------------------------------------
## Proportions:  0.642  vs  0.564
## Difference:  0.077
## Chi-squared:  5.401
## Degrees of freedom:  1
## P-value:  0.02012
## 95% CI: [ 0.011 ,  0.144 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o3-Low
## ----------------------------------------
## Proportions:  0.642  vs  0.539
## Difference:  0.102
## Chi-squared:  9.513
## Degrees of freedom:  1
## P-value:  0.00204
## 95% CI: [ 0.035 ,  0.169 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs GPT-5-High
## ----------------------------------------
## Proportions:  0.642  vs  0.67
## Difference:  -0.028
## Chi-squared:  1.138
## Degrees of freedom:  1
## P-value:  0.286
## 95% CI: [ -0.079 ,  0.022 ]
## Significant:  NO
##
##  o3-High vs o3-Pro
## ----------------------------------------
## Proportions:  0.642  vs  0.666
## Difference:  -0.025
## Chi-squared:  1.106
## Degrees of freedom:  1
## P-value:  0.2928
## 95% CI: [ -0.07 ,  0.02 ]
## Significant:  NO
##
##  o3-High vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.642  vs  0.595
## Difference:  0.047
## Chi-squared:  1.924
## Degrees of freedom:  1
## P-value:  0.1654
## 95% CI: [ -0.019 ,  0.113 ]
## Significant:  NO
##
##  o3-High vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.642  vs  0.506
## Difference:  0.136
## Chi-squared:  16.897
## Degrees of freedom:  1
## P-value:  0.00003946
## 95% CI: [ 0.069 ,  0.203 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.642  vs  0.407
## Difference:  0.234
## Chi-squared:  49.803
## Degrees of freedom:  1
## P-value:  0.0000000000017
## 95% CI: [ 0.168 ,  0.3 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o4-mini-High
## ----------------------------------------
## Proportions:  0.642  vs  0.53
## Difference:  0.111
## Chi-squared:  18.085
## Degrees of freedom:  1
## P-value:  0.00002112
## 95% CI: [ 0.059 ,  0.163 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.642  vs  0.49
## Difference:  0.152
## Chi-squared:  33.555
## Degrees of freedom:  1
## P-value:  0.000000006929
## 95% CI: [ 0.1 ,  0.204 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.642  vs  0.553
## Difference:  0.088
## Chi-squared:  16.139
## Degrees of freedom:  1
## P-value:  0.00005886
## 95% CI: [ 0.045 ,  0.131 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.642  vs  0.549
## Difference:  0.093
## Chi-squared:  7.863
## Degrees of freedom:  1
## P-value:  0.005045
## 95% CI: [ 0.026 ,  0.16 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs o3-Low
## ----------------------------------------
## Proportions:  0.564  vs  0.539
## Difference:  0.025
## Chi-squared:  0.282
## Degrees of freedom:  1
## P-value:  0.5956
## 95% CI: [ -0.058 ,  0.108 ]
## Significant:  NO
##
##  o3-Medium vs GPT-5-High
## ----------------------------------------
## Proportions:  0.564  vs  0.67
## Difference:  -0.106
## Chi-squared:  9.15
## Degrees of freedom:  1
## P-value:  0.002487
## 95% CI: [ -0.176 ,  -0.035 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs o3-Pro
## ----------------------------------------
## Proportions:  0.564  vs  0.666
## Difference:  -0.102
## Chi-squared:  9.74
## Degrees of freedom:  1
## P-value:  0.001803
## 95% CI: [ -0.168 ,  -0.036 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.564  vs  0.595
## Difference:  -0.03
## Chi-squared:  0.453
## Degrees of freedom:  1
## P-value:  0.5009
## 95% CI: [ -0.113 ,  0.052 ]
## Significant:  NO
```

```
##
##  o3-Medium vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.564  vs  0.506
## Difference:  0.059
## Chi-squared:  1.848
## Degrees of freedom:  1
## P-value:  0.174
## 95% CI: [ -0.024 ,  0.142 ]
## Significant:  NO
##
##  o3-Medium vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.564  vs  0.407
## Difference:  0.157
## Chi-squared:  14.152
## Degrees of freedom:  1
## P-value:  0.0001686
## 95% CI: [ 0.075 ,  0.239 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs o4-mini-High
## ----------------------------------------
## Proportions:  0.564  vs  0.53
## Difference:  0.034
## Chi-squared:  0.799
## Degrees of freedom:  1
## P-value:  0.3715
## 95% CI: [ -0.037 ,  0.105 ]
## Significant:  NO
##
##  o3-Medium vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.564  vs  0.49
## Difference:  0.075
## Chi-squared:  4.173
## Degrees of freedom:  1
## P-value:  0.04108
## 95% CI: [ 0.003 ,  0.146 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.564  vs  0.553
## Difference:  0.011
## Chi-squared:  0.072
## Degrees of freedom:  1
## P-value:  0.7878
## 95% CI: [ -0.054 ,  0.076 ]
## Significant:  NO
##
##  o3-Medium vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.564  vs  0.549
## Difference:  0.016
## Chi-squared:  0.093
## Degrees of freedom:  1
## P-value:  0.7605
## 95% CI: [ -0.067 ,  0.099 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-High
## ----------------------------------------
## Proportions:  0.539  vs  0.67
## Difference:  -0.13
## Chi-squared:  13.975
## Degrees of freedom:  1
## P-value:  0.0001853
## 95% CI: [ -0.201 ,  -0.06 ]
## Significant:  YES (p < 0.05)
##
##  o3-Low vs o3-Pro
## ----------------------------------------
## Proportions:  0.539  vs  0.666
## Difference:  -0.127
## Chi-squared:  15.089
## Degrees of freedom:  1
## P-value:  0.0001026
## 95% CI: [ -0.193 ,  -0.06 ]
## Significant:  YES (p < 0.05)
##
##  o3-Low vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.539  vs  0.595
## Difference:  -0.055
## Chi-squared:  1.653
## Degrees of freedom:  1
## P-value:  0.1985
## 95% CI: [ -0.138 ,  0.027 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.539  vs  0.506
## Difference:  0.034
## Chi-squared:  0.558
## Degrees of freedom:  1
## P-value:  0.4549
## 95% CI: [ -0.049 ,  0.117 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.539  vs  0.407
## Difference:  0.132
## Chi-squared:  9.956
## Degrees of freedom:  1
## P-value:  0.001603
## 95% CI: [ 0.049 ,  0.215 ]
## Significant:  YES (p < 0.05)
##
##  o3-Low vs o4-mini-High
## ----------------------------------------
## Proportions:  0.539  vs  0.53
## Difference:  0.009
## Chi-squared:  0.035
## Degrees of freedom:  1
## P-value:  0.8516
## 95% CI: [ -0.063 ,  0.081 ]
## Significant:  NO
##
##  o3-Low vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.539  vs  0.49
## Difference:  0.05
## Chi-squared:  1.791
## Degrees of freedom:  1
## P-value:  0.1808
## 95% CI: [ -0.022 ,  0.121 ]
## Significant:  NO
##
##  o3-Low vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.539  vs  0.553
## Difference:  -0.014
## Chi-squared:  0.142
## Degrees of freedom:  1
## P-value:  0.7067
## 95% CI: [ -0.079 ,  0.051 ]
## Significant:  NO
##
##  o3-Low vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.539  vs  0.549
## Difference:  -0.009
## Chi-squared:  0.021
## Degrees of freedom:  1
## P-value:  0.8856
## 95% CI: [ -0.092 ,  0.074 ]
## Significant:  NO
##
##  GPT-5-High vs o3-Pro
## ----------------------------------------
## Proportions:  0.67  vs  0.666
## Difference:  0.003
## Chi-squared:  0.007
## Degrees of freedom:  1
## P-value:  0.9336
## 95% CI: [ -0.047 ,  0.053 ]
## Significant:  NO
##
##  GPT-5-High vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.67  vs  0.595
## Difference:  0.075
## Chi-squared:  4.594
## Degrees of freedom:  1
## P-value:  0.03208
## 95% CI: [ 0.005 ,  0.145 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.67  vs  0.506
## Difference:  0.164
## Chi-squared:  22.084
## Degrees of freedom:  1
## P-value:  0.000002609
## 95% CI: [ 0.094 ,  0.235 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.67  vs  0.407
## Difference:  0.262
## Chi-squared:  55.522
## Degrees of freedom:  1
## P-value:  0.0000000000000924
## 95% CI: [ 0.193 ,  0.332 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o4-mini-High
## ----------------------------------------
## Proportions:  0.67  vs  0.53
## Difference:  0.139
## Chi-squared:  23.742
## Degrees of freedom:  1
## P-value:  0.000001102
## 95% CI: [ 0.083 ,  0.196 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.67  vs  0.49
## Difference:  0.18
## Chi-squared:  39.241
## Degrees of freedom:  1
## P-value:  0.0000000003745
## 95% CI: [ 0.124 ,  0.237 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.67  vs  0.553
## Difference:  0.116
## Chi-squared:  21.894
## Degrees of freedom:  1
## P-value:  0.000002882
## 95% CI: [ 0.068 ,  0.164 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.67  vs  0.549
## Difference:  0.121
## Chi-squared:  12.08
## Degrees of freedom:  1
## P-value:  0.0005097
## 95% CI: [ 0.051 ,  0.191 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.666  vs  0.595
## Difference:  0.072
## Chi-squared:  4.747
## Degrees of freedom:  1
## P-value:  0.02934
## 95% CI: [ 0.006 ,  0.137 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.666  vs  0.506
## Difference:  0.161
## Chi-squared:  24.15
## Degrees of freedom:  1
## P-value:  0.0000008913
## 95% CI: [ 0.094 ,  0.227 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.666  vs  0.407
## Difference:  0.259
## Chi-squared:  61.843
## Degrees of freedom:  1
## P-value:  0.000000000000003719
## 95% CI: [ 0.193 ,  0.325 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o4-mini-High
## ----------------------------------------
## Proportions:  0.666  vs  0.53
## Difference:  0.136
## Chi-squared:  27.478
## Degrees of freedom:  1
## P-value:  0.0000001589
## 95% CI: [ 0.084 ,  0.188 ]
## Significant:  YES (p < 0.05)
```

```
##
##  o3-Pro vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.666  vs  0.49
## Difference:  0.177
## Chi-squared:  45.953
## Degrees of freedom:  1
## P-value:  0.00000000001211
## 95% CI: [ 0.125 ,  0.229 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.666  vs  0.553
## Difference:  0.113
## Chi-squared:  26.82
## Degrees of freedom:  1
## P-value:  0.0000002234
## 95% CI: [ 0.07 ,  0.155 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.666  vs  0.549
## Difference:  0.118
## Chi-squared:  12.982
## Degrees of freedom:  1
## P-value:  0.0003144
## 95% CI: [ 0.051 ,  0.184 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.595  vs  0.506
## Difference:  0.089
## Chi-squared:  4.464
## Degrees of freedom:  1
## P-value:  0.03461
## 95% CI: [ 0.007 ,  0.172 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.595  vs  0.407
## Difference:  0.187
## Chi-squared:  20.31
## Degrees of freedom:  1
## P-value:  0.000006585
## 95% CI: [ 0.105 ,  0.269 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs o4-mini-High
## ----------------------------------------
## Proportions:  0.595  vs  0.53
## Difference:  0.064
## Chi-squared:  3.103
## Degrees of freedom:  1
## P-value:  0.07814
## 95% CI: [ -0.006 ,  0.135 ]
## Significant:  NO
##
##  GPT-5-Medium vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.595  vs  0.49
## Difference:  0.105
## Chi-squared:  8.451
## Degrees of freedom:  1
## P-value:  0.003648
## 95% CI: [ 0.034 ,  0.176 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.595  vs  0.553
## Difference:  0.041
## Chi-squared:  1.49
## Degrees of freedom:  1
## P-value:  0.2222
## 95% CI: [ -0.023 ,  0.106 ]
## Significant:  NO
##
##  GPT-5-Medium vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.595  vs  0.549
## Difference:  0.046
## Chi-squared:  1.124
## Degrees of freedom:  1
## P-value:  0.2891
## 95% CI: [ -0.036 ,  0.129 ]
## Significant:  NO
##
##  GPT-5-Low vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.506  vs  0.407
## Difference:  0.098
## Chi-squared:  5.436
## Degrees of freedom:  1
## P-value:  0.01972
## 95% CI: [ 0.016 ,  0.181 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Low vs o4-mini-High
## ----------------------------------------
## Proportions:  0.506  vs  0.53
## Difference:  -0.025
## Chi-squared:  0.395
## Degrees of freedom:  1
## P-value:  0.5295
## 95% CI: [ -0.096 ,  0.047 ]
## Significant:  NO
##
##  GPT-5-Low vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.506  vs  0.49
## Difference:  0.016
## Chi-squared:  0.146
## Degrees of freedom:  1
## P-value:  0.7026
## 95% CI: [ -0.056 ,  0.088 ]
## Significant:  NO
##
##  GPT-5-Low vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.506  vs  0.553
## Difference:  -0.048
## Chi-squared:  2.038
## Degrees of freedom:  1
## P-value:  0.1534
## 95% CI: [ -0.113 ,  0.017 ]
## Significant:  NO
##
##  GPT-5-Low vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.506  vs  0.549
## Difference:  -0.043
## Chi-squared:  0.947
## Degrees of freedom:  1
## P-value:  0.3306
## 95% CI: [ -0.126 ,  0.04 ]
## Significant:  NO
##
##  GPT-5-Minimal vs o4-mini-High
## ----------------------------------------
## Proportions:  0.407  vs  0.53
## Difference:  -0.123
## Chi-squared:  11.596
## Degrees of freedom:  1
## P-value:  0.0006608
## 95% CI: [ -0.194 ,  -0.052 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Minimal vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.407  vs  0.49
## Difference:  -0.082
## Chi-squared:  5.106
## Degrees of freedom:  1
## P-value:  0.02384
## 95% CI: [ -0.153 ,  -0.011 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Minimal vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.407  vs  0.553
## Difference:  -0.146
## Chi-squared:  19.968
## Degrees of freedom:  1
## P-value:  0.000007874
## 95% CI: [ -0.211 ,  -0.082 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Minimal vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.407  vs  0.549
## Difference:  -0.141
## Chi-squared:  11.419
## Degrees of freedom:  1
## P-value:  0.0007269
## 95% CI: [ -0.224 ,  -0.059 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-High vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.53  vs  0.49
## Difference:  0.041
## Chi-squared:  1.83
## Degrees of freedom:  1
## P-value:  0.1761
## 95% CI: [ -0.017 ,  0.099 ]
## Significant:  NO
##
##  o4-mini-High vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.53  vs  0.553
## Difference:  -0.023
## Chi-squared:  0.782
## Degrees of freedom:  1
## P-value:  0.3765
## 95% CI: [ -0.073 ,  0.027 ]
## Significant:  NO
##
##  o4-mini-High vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.53  vs  0.549
## Difference:  -0.018
## Chi-squared:  0.2
## Degrees of freedom:  1
## P-value:  0.6544
## 95% CI: [ -0.09 ,  0.053 ]
## Significant:  NO
##
##  o4-mini-Medium vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.49  vs  0.553
## Difference:  -0.064
## Chi-squared:  6.321
## Degrees of freedom:  1
## P-value:  0.01193
## 95% CI: [ -0.114 ,  -0.014 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-Medium vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.49  vs  0.549
## Difference:  -0.059
## Chi-squared:  2.554
## Degrees of freedom:  1
## P-value:  0.11
## 95% CI: [ -0.131 ,  0.013 ]
## Significant:  NO
##
##  o3-GPT-Image-High vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.553  vs  0.549
## Difference:  0.005
## Chi-squared:  0.008
## Degrees of freedom:  1
## P-value:  0.9281
## 95% CI: [ -0.06 ,  0.07 ]
## Significant:  NO
```

```
# Summary table
combined_reasoning_summary <- combined_reasoning_results %>%
  select(comparison, diff, chi_squared, p_value, significant) %>%
  mutate(diff = round(diff, 3),
         p_value = round(p_value, 4))
cat("\n\nSummary Table - Combined Reasoning Variations:\n")
```

```
##
##
## Summary Table - Combined Reasoning Variations:
```

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

```
##
##
##               comparison                                   diff   chi_squared   p_value  significan
## ------------  ----------------------------------------  -------  ------------  --------  ----------
## X-squared     Humans vs o3-High                          -0.094    28.6308876    0.0000  TRUE
## X-squared1    Humans vs o3-Medium                        -0.017     0.2729579    0.6014  FALSE
## X-squared2    Humans vs o3-Low                            0.008     0.0434201    0.8349  FALSE
## X-squared3    Humans vs GPT-5-High                       -0.123    33.3019857    0.0000  TRUE
## X-squared4    Humans vs o3-Pro                           -0.119    45.7603911    0.0000  TRUE
## X-squared5    Humans vs GPT-5-Medium                     -0.048     2.4410000    0.1182  FALSE
## X-squared6    Humans vs GPT-5-Low                         0.042     1.8538008    0.1733  FALSE
## X-squared7    Humans vs GPT-5-Minimal                     0.140    22.1514893    0.0000  TRUE
## X-squared8    Humans vs o4-mini-High                      0.017     0.5765459    0.4477  FALSE
## X-squared9    Humans vs o4-mini-Medium                    0.058     7.2087261    0.0073  TRUE
## X-squared10   Humans vs o3-GPT-Image-High                -0.006     0.1425741    0.7057  FALSE
## X-squared11   Humans vs o3-GPT-Image-Medium              -0.001     0.0000000    1.0000  FALSE
## X-squared12   o3-High vs o3-Medium                        0.077     5.4014481    0.0201  TRUE
## X-squared13   o3-High vs o3-Low                            0.102     9.5128679    0.0020  TRUE
## X-squared14   o3-High vs GPT-5-High                       -0.028     1.1383107    0.2860  FALSE
## X-squared15   o3-High vs o3-Pro                           -0.025     1.1064986    0.2928  FALSE
## X-squared16   o3-High vs GPT-5-Medium                      0.047     1.9244960    0.1654  FALSE
## X-squared17   o3-High vs GPT-5-Low                         0.136    16.8973639    0.0000  TRUE
## X-squared18   o3-High vs GPT-5-Minimal                     0.234    49.8027874    0.0000  TRUE
## X-squared19   o3-High vs o4-mini-High                      0.111    18.0851641    0.0000  TRUE
## X-squared20   o3-High vs o4-mini-Medium                    0.152    33.5545083    0.0000  TRUE
## X-squared21   o3-High vs o3-GPT-Image-High                 0.088    16.1391378    0.0001  TRUE
## X-squared22   o3-High vs o3-GPT-Image-Medium               0.093     7.8633176    0.0050  TRUE
## X-squared23   o3-Medium vs o3-Low                          0.025     0.2817014    0.5956  FALSE
## X-squared24   o3-Medium vs GPT-5-High                     -0.106     9.1499944    0.0025  TRUE
## X-squared25   o3-Medium vs o3-Pro                         -0.102     9.7395028    0.0018  TRUE
## X-squared26   o3-Medium vs GPT-5-Medium                   -0.030     0.4530888    0.5009  FALSE
## X-squared27   o3-Medium vs GPT-5-Low                       0.059     1.8478915    0.1740  FALSE
## X-squared28   o3-Medium vs GPT-5-Minimal                   0.157    14.1524588    0.0002  TRUE
## X-squared29   o3-Medium vs o4-mini-High                    0.034     0.7985134    0.3715  FALSE
## X-squared30   o3-Medium vs o4-mini-Medium                  0.075     4.1728658    0.0411  TRUE
## X-squared31   o3-Medium vs o3-GPT-Image-High               0.011     0.0724599    0.7878  FALSE
## X-squared32   o3-Medium vs o3-GPT-Image-Medium             0.016     0.0929174    0.7605  FALSE
## X-squared33   o3-Low vs GPT-5-High                        -0.130    13.9749251    0.0002  TRUE
## X-squared34   o3-Low vs o3-Pro                            -0.127    15.0885546    0.0001  TRUE
## X-squared35   o3-Low vs GPT-5-Medium                      -0.055     1.6530660    0.1985  FALSE
## X-squared36   o3-Low vs GPT-5-Low                          0.034     0.5584009    0.4549  FALSE
## X-squared37   o3-Low vs GPT-5-Minimal                      0.132     9.9564864    0.0016  TRUE
## X-squared38   o3-Low vs o4-mini-High                       0.009     0.0349944    0.8516  FALSE
## X-squared39   o3-Low vs o4-mini-Medium                     0.050     1.7906703    0.1808  FALSE
## X-squared40   o3-Low vs o3-GPT-Image-High                 -0.014     0.1416433    0.7067  FALSE
## X-squared41   o3-Low vs o3-GPT-Image-Medium               -0.009     0.0207156    0.8856  FALSE
## X-squared42   GPT-5-High vs o3-Pro                         0.003     0.0069479    0.9336  FALSE
## X-squared43   GPT-5-High vs GPT-5-Medium                   0.075     4.5940970    0.0321  TRUE
## X-squared44   GPT-5-High vs GPT-5-Low                      0.164    22.0842822    0.0000  TRUE
## X-squared45   GPT-5-High vs GPT-5-Minimal                  0.262    55.5223070    0.0000  TRUE
## X-squared46   GPT-5-High vs o4-mini-High                   0.139    23.7419837    0.0000  TRUE
## X-squared47   GPT-5-High vs o4-mini-Medium                 0.180    39.2413370    0.0000  TRUE
## X-squared48   GPT-5-High vs o3-GPT-Image-High              0.116    21.8935396    0.0000  TRUE
## X-squared49   GPT-5-High vs o3-GPT-Image-Medium            0.121    12.0796975    0.0005  TRUE
## X-squared50   o3-Pro vs GPT-5-Medium                       0.072     4.7472865    0.0293  TRUE
## X-squared51   o3-Pro vs GPT-5-Low                          0.161    24.1496509    0.0000  TRUE
## X-squared52   o3-Pro vs GPT-5-Minimal                      0.259    61.8432618    0.0000  TRUE
## X-squared53   o3-Pro vs o4-mini-High                       0.136    27.4784870    0.0000  TRUE
## X-squared54   o3-Pro vs o4-mini-Medium                     0.177    45.9534117    0.0000  TRUE
## X-squared55   o3-Pro vs o3-GPT-Image-High                  0.113    26.8195138    0.0000  TRUE
## X-squared56   o3-Pro vs o3-GPT-Image-Medium                0.118    12.9823165    0.0003  TRUE
## X-squared57   GPT-5-Medium vs GPT-5-Low                    0.089     4.4643993    0.0346  TRUE
## X-squared58   GPT-5-Medium vs GPT-5-Minimal                0.187    20.3101779    0.0000  TRUE
## X-squared59   GPT-5-Medium vs o4-mini-High                 0.064     3.1032648    0.0781  FALSE
## X-squared60   GPT-5-Medium vs o4-mini-Medium               0.105     8.4511203    0.0036  TRUE
## X-squared61   GPT-5-Medium vs o3-GPT-Image-High            0.041     1.4902319    0.2222  FALSE
## X-squared62   GPT-5-Medium vs o3-GPT-Image-Medium          0.046     1.1236450    0.2891  FALSE
## X-squared63   GPT-5-Low vs GPT-5-Minimal                   0.098     5.4363969    0.0197  TRUE
## X-squared64   GPT-5-Low vs o4-mini-High                   -0.025     0.3954028    0.5295  FALSE
## X-squared65   GPT-5-Low vs o4-mini-Medium                  0.016     0.1457727    0.7026  FALSE
## X-squared66   GPT-5-Low vs o3-GPT-Image-High              -0.048     2.0376965    0.1534  FALSE
## X-squared67   GPT-5-Low vs o3-GPT-Image-Medium          -0.043     0.9466394    0.3306  FALSE
## X-squared68   GPT-5-Minimal vs o4-mini-High             -0.123    11.5963146    0.0007  TRUE
## X-squared69   GPT-5-Minimal vs o4-mini-Medium           -0.082     5.1059649    0.0238  TRUE
## X-squared70   GPT-5-Minimal vs o3-GPT-Image-High        -0.146    19.9681005    0.0000  TRUE
## X-squared71   GPT-5-Minimal vs o3-GPT-Image-Medium      -0.141    11.4191029    0.0007  TRUE
## X-squared72   o4-mini-High vs o4-mini-Medium             0.041     1.8298142    0.1761  FALSE
## X-squared73   o4-mini-High vs o3-GPT-Image-High         -0.023     0.7821301    0.3765  FALSE
## X-squared74   o4-mini-High vs o3-GPT-Image-Medium       -0.018     0.2004081    0.6544  FALSE
## X-squared75   o4-mini-Medium vs o3-GPT-Image-High       -0.064     6.3212724    0.0119  TRUE
## X-squared76   o4-mini-Medium vs o3-GPT-Image-Medium     -0.059     2.5541219    0.1100  FALSE
## X-squared77   o3-GPT-Image-High vs o3-GPT-Image-Medium   0.005     0.0081513    0.9281  FALSE
```

```
# Count significant differences
combined_reasoning_sig_count <- sum(combined_reasoning_results$significant)
cat("\n\nCombined Reasoning Variations Summary:\n")
```

```
##
##
## Combined Reasoning Variations Summary:
```

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

```
##   Total comparisons: 78
```

```
cat("  Significant differences:", combined_reasoning_sig_count, "\n")
```

```
##   Significant differences: 43
```

```
cat("  Percentage significant:", round(combined_reasoning_sig_count / nrow(combined_reasoning_results)
```

```
##   Percentage significant: 55.1 %
```

```
# Show significant comparisons
cat("Significant Comparisons in Combined Reasoning Variations:\n")
```

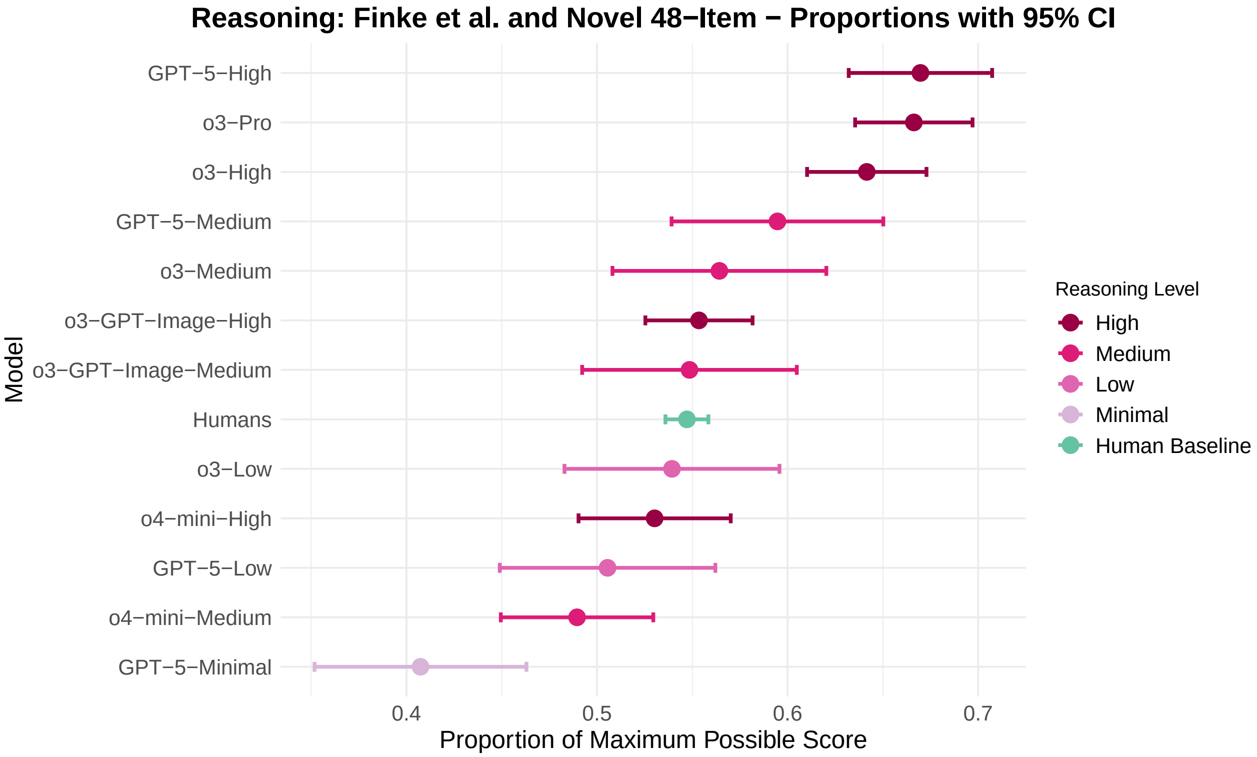
```
## Significant Comparisons in Combined Reasoning Variations:
```

```
combined_reasoning_sig <- combined_reasoning_results[combined_reasoning_results$significant, c("compari
if (nrow(combined_reasoning_sig) > 0) {
  print(kable(combined_reasoning_sig, format = "simple", digits = 4))
} else {
  cat("  None\n")
}
```

```
##
##
##               comparison                               diff   p_value
## ------------  -------------------------------------  --------  --------
## X-squared     Humans vs o3-High                       -0.0944    0.0000
## X-squared3    Humans vs GPT-5-High                    -0.1225    0.0000
## X-squared4    Humans vs o3-Pro                        -0.1191    0.0000
## X-squared7    Humans vs GPT-5-Minimal                  0.1398    0.0000
## X-squared9    Humans vs o4-mini-Medium                 0.0576    0.0073
## X-squared12   o3-High vs o3-Medium                     0.0773    0.0201
## X-squared13   o3-High vs o3-Low                        0.1022    0.0020
## X-squared17   o3-High vs GPT-5-Low                     0.1360    0.0000
## X-squared18   o3-High vs GPT-5-Minimal                 0.2342    0.0000
## X-squared19   o3-High vs o4-mini-High                  0.1113    0.0000
## X-squared20   o3-High vs o4-mini-Medium                0.1520    0.0000
## X-squared21   o3-High vs o3-GPT-Image-High             0.0881    0.0001
## X-squared22   o3-High vs o3-GPT-Image-Medium           0.0930    0.0050
## X-squared24   o3-Medium vs GPT-5-High                 -0.1055    0.0025
## X-squared25   o3-Medium vs o3-Pro                     -0.1020    0.0018
## X-squared28   o3-Medium vs GPT-5-Minimal               0.1569    0.0002
## X-squared30   o3-Medium vs o4-mini-Medium              0.0747    0.0411
## X-squared33   o3-Low vs GPT-5-High                    -0.1304    0.0002
## X-squared34   o3-Low vs o3-Pro                        -0.1269    0.0001
## X-squared37   o3-Low vs GPT-5-Minimal                  0.1320    0.0016
## X-squared43   GPT-5-High vs GPT-5-Medium               0.0750    0.0321
## X-squared44   GPT-5-High vs GPT-5-Low                  0.1642    0.0000
## X-squared45   GPT-5-High vs GPT-5-Minimal              0.2624    0.0000
## X-squared46   GPT-5-High vs o4-mini-High               0.1395    0.0000
## X-squared47   GPT-5-High vs o4-mini-Medium             0.1802    0.0000
## X-squared48   GPT-5-High vs o3-GPT-Image-High          0.1162    0.0000
## X-squared49   GPT-5-High vs o3-GPT-Image-Medium        0.1212    0.0005
## X-squared50   o3-Pro vs GPT-5-Medium                   0.0716    0.0293
## X-squared51   o3-Pro vs GPT-5-Low                      0.1607    0.0000
## X-squared52   o3-Pro vs GPT-5-Minimal                  0.2589    0.0000
## X-squared53   o3-Pro vs o4-mini-High                   0.1360    0.0000
## X-squared54   o3-Pro vs o4-mini-Medium                 0.1767    0.0000
## X-squared55   o3-Pro vs o3-GPT-Image-High              0.1128    0.0000
## X-squared56   o3-Pro vs o3-GPT-Image-Medium            0.1177    0.0003
## X-squared57   GPT-5-Medium vs GPT-5-Low                0.0892    0.0346
## X-squared58   GPT-5-Medium vs GPT-5-Minimal            0.1873    0.0000
## X-squared60   GPT-5-Medium vs o4-mini-Medium           0.1052    0.0036
## X-squared63   GPT-5-Low vs GPT-5-Minimal               0.0982    0.0197
## X-squared68   GPT-5-Minimal vs o4-mini-High           -0.1229    0.0007
## X-squared69   GPT-5-Minimal vs o4-mini-Medium         -0.0822    0.0238
## X-squared70   GPT-5-Minimal vs o3-GPT-Image-High      -0.1461    0.0000
## X-squared71   GPT-5-Minimal vs o3-GPT-Image-Medium    -0.1412    0.0007
## X-squared75   o4-mini-Medium vs o3-GPT-Image-High     -0.0640    0.0119
```

Visualization of Combined Reasoning Variations

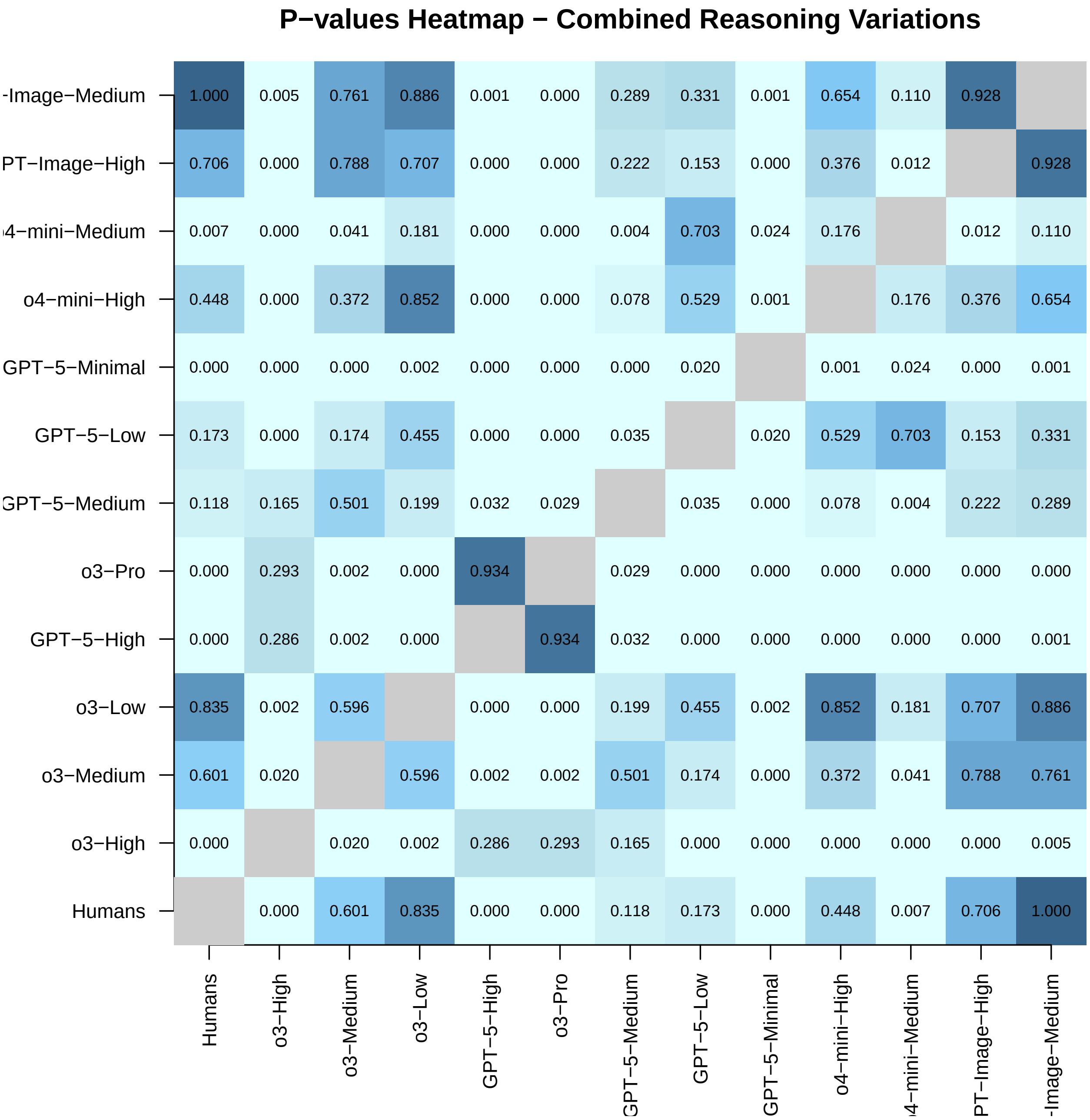
```r
# Plot proportions with confidence intervals for combined reasoning variations
combined_reasoning_plot <- ggplot(collapsed_reasoning_data, aes(x = reorder(model, proportion), y = prop
  geom_point(size = 4, aes(color = color)) +
  geom_errorbar(aes(ymin = proportion - 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    ymax = proportion + 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    color = color),
                width = 0.2, size = 1) +
  coord_flip() +
  theme_minimal() +
  labs(title = "Reasoning: Finke et al. and Novel 48-Item - Proportions with 95% CI",
       x = "Model",
       y = "Proportion of Maximum Possible Score") +
  theme(plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
        axis.text = element_text(size = 12),
        axis.title = element_text(size = 14),
        legend.text = element_text(size = 12)) +
  scale_color_manual(
    values = c("#980043", "#dd1c77", "#df65b0", "#d7b5d8", "#66c2a5"),
    name = "Reasoning Level",
    breaks = c("#980043", "#dd1c77", "#df65b0", "#d7b5d8", "#66c2a5"),
    labels = c("High", "Medium", "Low", "Minimal", "Human Baseline")
  )

# minimal #d7b5d8
# low #df65b0
# medium #dd1c77
# high #980043
# human #66c2a5



print(combined_reasoning_plot)
```



Heatmap for Combined Reasoning Variations

```
# Create matrix of p-values for combined reasoning variations
combined_reasoning_models <- collapsed_reasoning_data$model
combined_reasoning_pval_matrix <- matrix(NA, nrow = length(combined_reasoning_models), ncol = length(com
rownames(combined_reasoning_pval_matrix) <- combined_reasoning_models
colnames(combined_reasoning_pval_matrix) <- combined_reasoning_models

for (i in 1:nrow(combined_reasoning_results)) {
  row_idx <- which(combined_reasoning_models == combined_reasoning_results$model1[i])
  col_idx <- which(combined_reasoning_models == combined_reasoning_results$model2[i])
  combined_reasoning_pval_matrix[row_idx, col_idx] <- combined_reasoning_results$p_value[i]
  combined_reasoning_pval_matrix[col_idx, row_idx] <- combined_reasoning_results$p_value[i]
}
# Set diagonal to NA
```

```
diag(combined_reasoning_pval_matrix) <- NA
# Set margins for better label display
par(mar = c(6, 6, 3, 2))
# Plot heatmap with same color palette
image(combined_reasoning_pval_matrix, axes = FALSE, col = col_palette,
      main = "P-values Heatmap - Combined Reasoning Variations")
axis(1, at = seq(0, 1, length.out = length(combined_reasoning_models)), labels = combined_reasoning_mode
     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(combined_reasoning_models)), labels = combined_reasoning_mode
     las = 2, cex.axis = 0.8)
# Add gray color for diagonal
for (i in 1:length(combined_reasoning_models)) {
  x_pos <- (i - 1) / (length(combined_reasoning_models) - 1)
  y_pos <- (i - 1) / (length(combined_reasoning_models) - 1)
  rect(x_pos - 0.5 / (length(combined_reasoning_models) - 1), y_pos - 0.5 / (length(combined_reasoning_m
       x_pos + 0.5 / (length(combined_reasoning_models) - 1), y_pos + 0.5 / (length(combined_reasoning_m
       col = "gray80", border = NA)
}
# Add p-values to the plot
for (i in 1:nrow(combined_reasoning_pval_matrix)) {
  for (j in 1:ncol(combined_reasoning_pval_matrix)) {
    if (!is.na(combined_reasoning_pval_matrix[i, j])) {
      x_pos <- (j - 1) / (ncol(combined_reasoning_pval_matrix) - 1)
      y_pos <- (i - 1) / (nrow(combined_reasoning_pval_matrix) - 1)
      text(x_pos, y_pos, sprintf("%.3f", combined_reasoning_pval_matrix[i, j]), cex = 0.7)
    }
  }
}
```



	Humans	o3−High	o3−Medium	o3−Low	GPT−5−High	o3−Pro	GPT−5−Medium	GPT−5−Low	GPT−5−Minimal	o4−mini−High	4−mini−Medium	PT−Image−High	-Image−Medium
-Image−Medium	1.000	0.005	0.761	0.886	0.001	0.000	0.289	0.331	0.001	0.654	0.110	0.928	
PT−Image−High	0.706	0.000	0.788	0.707	0.000	0.000	0.222	0.153	0.000	0.376	0.012		0.928
4−mini−Medium	0.007	0.000	0.041	0.181	0.000	0.000	0.004	0.703	0.024	0.176		0.012	0.110
o4−mini−High	0.448	0.000	0.372	0.852	0.000	0.000	0.078	0.529	0.001		0.176	0.376	0.654
GPT−5−Minimal	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.020		0.001	0.024	0.000	0.001
GPT−5−Low	0.173	0.000	0.174	0.455	0.000	0.000	0.035		0.020	0.529	0.703	0.153	0.331
GPT−5−Medium	0.118	0.165	0.501	0.199	0.032	0.029		0.035	0.000	0.078	0.004	0.222	0.289
o3−Pro	0.000	0.293	0.002	0.000	0.934		0.029	0.000	0.000	0.000	0.000	0.000	0.000
GPT−5−High	0.000	0.286	0.002	0.000		0.934	0.032	0.000	0.000	0.000	0.000	0.000	0.001
o3−Low	0.835	0.002	0.596		0.000	0.000	0.199	0.455	0.002	0.852	0.181	0.707	0.886
o3−Medium	0.601	0.020		0.596	0.002	0.002	0.501	0.174	0.000	0.372	0.041	0.788	0.761
o3−High	0.000		0.020	0.002	0.286	0.293	0.165	0.000	0.000	0.000	0.000	0.000	0.005
Humans		0.000	0.601	0.835	0.000	0.000	0.118	0.173	0.000	0.448	0.007	0.706	1.000

Summary of Significant Differences - Combined Reasoning Variations

```r
# Count significant differences for combined reasoning variations
combined_reasoning_sig_count <- sum(combined_reasoning_results$significant)
cat("Summary of Significant Differences - Combined Reasoning Variations:\n")
```

```
## Summary of Significant Differences - Combined Reasoning Variations:
```

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

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

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

```
##   Total comparisons: 78
```

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

```
##   Significant differences: 43
```

```r
cat("  Percentage significant:", round(combined_reasoning_sig_count / nrow(combined_reasoning_results)
```

```
##   Percentage significant: 55.1 %
```

```r
# Show which comparisons are significant
cat("Significant Comparisons in Combined Reasoning Variations:\n")
```

```
## Significant Comparisons in Combined Reasoning Variations:
```

```r
combined_reasoning_sig <- combined_reasoning_results[combined_reasoning_results$significant, c("compari
if (nrow(combined_reasoning_sig) > 0) {
  print(kable(combined_reasoning_sig, format = "simple", digits = 4))
} else {
  cat("  None\n")
}
```

```
## 
## 
##               comparison                                  diff    p_value
## ------------  -------------------------------------  ---------  ---------
## X-squared     Humans vs o3-High                        -0.0944     0.0000
## X-squared3    Humans vs GPT-5-High                     -0.1225     0.0000
## X-squared4    Humans vs o3-Pro                         -0.1191     0.0000
## X-squared7    Humans vs GPT-5-Minimal                   0.1398     0.0000
## X-squared9    Humans vs o4-mini-Medium                  0.0576     0.0073
## X-squared12   o3-High vs o3-Medium                      0.0773     0.0201
## X-squared13   o3-High vs o3-Low                         0.1022     0.0020
## X-squared17   o3-High vs GPT-5-Low                      0.1360     0.0000
## X-squared18   o3-High vs GPT-5-Minimal                  0.2342     0.0000
## X-squared19   o3-High vs o4-mini-High                   0.1113     0.0000
## X-squared20   o3-High vs o4-mini-Medium                 0.1520     0.0000
## X-squared21   o3-High vs o3-GPT-Image-High              0.0881     0.0001
## X-squared22   o3-High vs o3-GPT-Image-Medium            0.0930     0.0050
## X-squared24   o3-Medium vs GPT-5-High                  -0.1055     0.0025
## X-squared25   o3-Medium vs o3-Pro                      -0.1020     0.0018
## X-squared28   o3-Medium vs GPT-5-Minimal                0.1569     0.0002
## X-squared30   o3-Medium vs o4-mini-Medium               0.0747     0.0411
## X-squared33   o3-Low vs GPT-5-High                     -0.1304     0.0002
## X-squared34   o3-Low vs o3-Pro                         -0.1269     0.0001
## X-squared37   o3-Low vs GPT-5-Minimal                   0.1320     0.0016
## X-squared43   GPT-5-High vs GPT-5-Medium                0.0750     0.0321
## X-squared44   GPT-5-High vs GPT-5-Low                   0.1642     0.0000
## X-squared45   GPT-5-High vs GPT-5-Minimal               0.2624     0.0000
## X-squared46   GPT-5-High vs o4-mini-High                0.1395     0.0000
## X-squared47   GPT-5-High vs o4-mini-Medium              0.1802     0.0000
## X-squared48   GPT-5-High vs o3-GPT-Image-High           0.1162     0.0000
## X-squared49   GPT-5-High vs o3-GPT-Image-Medium         0.1212     0.0005
## X-squared50   o3-Pro vs GPT-5-Medium                    0.0716     0.0293
## X-squared51   o3-Pro vs GPT-5-Low                       0.1607     0.0000
## X-squared52   o3-Pro vs GPT-5-Minimal                   0.2589     0.0000
## X-squared53   o3-Pro vs o4-mini-High                    0.1360     0.0000
## X-squared54   o3-Pro vs o4-mini-Medium              0.1767    0.0000
## X-squared55   o3-Pro vs o3-GPT-Image-High           0.1128    0.0000
## X-squared56   o3-Pro vs o3-GPT-Image-Medium         0.1177    0.0003
## X-squared57   GPT-5-Medium vs GPT-5-Low             0.0892    0.0346
## X-squared58   GPT-5-Medium vs GPT-5-Minimal         0.1873    0.0000
## X-squared60   GPT-5-Medium vs o4-mini-Medium        0.1052    0.0036
## X-squared63   GPT-5-Low vs GPT-5-Minimal            0.0982    0.0197
## X-squared68   GPT-5-Minimal vs o4-mini-High        -0.1229    0.0007
## X-squared69   GPT-5-Minimal vs o4-mini-Medium      -0.0822    0.0238
## X-squared70   GPT-5-Minimal vs o3-GPT-Image-High   -0.1461    0.0000
## X-squared71   GPT-5-Minimal vs o3-GPT-Image-Medium -0.1412    0.0007
## X-squared75   o4-mini-Medium vs o3-GPT-Image-High  -0.0640    0.0119
```

Export Results to CSV

```r
# Combine all results
all_results <- rbind(finke_results, novel_48_results, collapsed_results,
                     finke_reasoning_results, novel_48_reasoning_results,
                     combined_reasoning_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'
```