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

Author

August 28, 2025

Summary

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


```
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$Model == "OpenAI: o4-mini - Multiple Context - High
  llm_data_finke[llm_data_finke$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_s
```

```r
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$Model == "OpenAI: GPT 5 - Multiple Context - High Reas
  llm_data_finke[llm_data_finke$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$Model == "DeepMind: Gemini 2.5 Pro - Multiple Con
  llm_data_finke[llm_data_finke$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 Con
opus4_1_finke_max_score <- (12) * 5

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

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

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

o3_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI: o3 - Single Context - High Reasoning
  llm_data_novel[llm_data_novel$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$Model == "OpenAI: o4-mini - Multiple Context - High
  llm_data_novel[llm_data_novel$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$Model == "OpenAI: GPT 4.1 - Multiple Context (2025-0
  llm_data_novel[llm_data_novel$Model == "OpenAI: GPT 4.1 - Single Context (2025-07-21)", "overall_score
gpt4_1_novel_max_score <- (48 + 48) * 5

gpt4_1_images_novel_score <- llm_data_novel[llm_data_novel$Model == "OpenAI: GPT 4.1 w/ GPT-image-1 - Mu
  llm_data_novel[llm_data_novel$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$Model == "OpenAI: GPT 5 - Multiple Context - High Reas
  llm_data_novel[llm_data_novel$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 Flash
gemini2_0_flash_images_novel_max_score <- (48) * 5

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
```

```r
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
```

```r
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"]) * 5

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"]) * 5


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

gpt_4_1_images_collapsed_mc <- llm_data_sc_mc[llm_data_sc_mc$Model == "gpt4.1_images_mc", "overall_score
gpt_4_1_images_collapsed_mc_max <- (llm_data_sc_mc[llm_data_sc_mc$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 +
  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
```

```
## 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
```

```r
gemini2_0_flash_images_total_score <- gemini2_0_flash_images_finke_score + gemini2_0_flash_images_novel
gemini2_0_flash_images_total_max_score <- gemini2_0_flash_images_finke_max_score + gemini2_0_flash_imag

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
```

```r
## 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
```

```r
# Create data frames for easier manipulation
finke_data <- data.frame(
  model = c("Humans", "o3", "o3-GPT-Image",
            "o3-Pro", "GPT-4.1", "GPT-4.1-GPT-Image",
            "ChatGPT-4o", "o4-mini", "Gemini-2.5",
            "Gemini-2.0-Flash", "Gemini-2.0-Flash-GPT-Image",
            "Sonnet-4", "Opus-4.1", "GPT-5"),
  score = c(humans_finke_score, o3_finke_score, o3_images_finke_score,
            o3_pro_finke_score, gpt4_1_finke_score, gpt4_1_images_finke_score,
            chatgpt_4o_finke_score, o4_mini_finke_score, gemini2_5_finke_score,
            gemini2_0_flash_finke_score, gemini2_0_flash_images_finke_score,
```

```
            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)
)

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

```
novel_data <- data.frame(
  model = c("Humans", "o3", "o3-GPT-Image",
            "o3-Pro", "GPT-4.1", "GPT-4.1-GPT-Image",
            "ChatGPT-4o", "o4-mini", "Gemini-2.5",
            "Gemini-2.0-Flash", "Gemini-2.0-Flash-GPT-Image",
            "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)

)

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

```
collapsed_data <- data.frame(
  model = c("Humans", "o3", "o3-GPT-Image",
            "o3-Pro", "GPT-4.1", "GPT-4.1-GPT-Image",
            "ChatGPT-4o", "o4-mini", "Gemini-2.5",
            "Gemini-2.0-Flash", "Gemini-2.0-Flash-GPT-Image",
            "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)
)

# 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',
            "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, 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, 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,
                o3_images_finke_max_score, medium_o3_images_finke_max_score)
)

# 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',
            "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_novel_score, o3_novel_score, medium_o3_novel_score,
            low_o3_novel_score,
            gpt5_novel_score, medium_gpt5_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, low_gpt5_novel_max_score,
                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)
)
# Calculate proportions from score/max_score
novel_reasoning_data$proportion <- novel_reasoning_data$score / novel_reasoning_data$max_score
```

```
collapsed_reasoning_data <- data.frame(
  model = c("Humans", "o3-High", "o3-Medium",
            "o3-Low", 'GPT-5-High',
            "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, 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, medium_gpt5_total_max_score, low_gpt5_total_max_score,
                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)
)
# Calculate proportions from score/max_score
collapsed_reasoning_data$proportion <- collapsed_reasoning_data$score / collapsed_reasoning_data$max_sc
```

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

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

```
print(finke_data)
```

```
##                         model     score max_score proportion
## 1                      Humans 952.09643      1525  0.6243255
## 2                          o3 109.15000       180  0.6063889
## 3                 o3-GPT-Image 136.23333       240  0.5676389
## 4                       o3-Pro 140.65833       180  0.7814352
## 5                     GPT-4.1  57.40714       120  0.4783929
## 6            GPT-4.1-GPT-Image  42.25000       120  0.3520833
## 7                   ChatGPT-4o  53.23095       120  0.4435913
## 8                      o4-mini  64.00833       120  0.5334028
## 9                    Gemini-2.5  62.87500       120  0.5239583
## 10            Gemini-2.0-Flash  41.60000       120  0.3466667
## 11 Gemini-2.0-Flash-GPT-Image  19.03810        60  0.3173016
## 12                    Sonnet-4  56.40238       120  0.4700198
## 13                    Opus-4.1  44.96667        60  0.7494444
## 14                       GPT-5  92.20000       120  0.7683333
```

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

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

```
print(novel_data)
```

```
##                         model      score max_score proportion
## 1                      Humans 3099.37024      5965  0.5195927
## 2                          o3  472.74048       720  0.6565840
## 3                 o3-GPT-Image  530.69881       960  0.5528113
## 4                       o3-Pro  457.21310       720  0.6350182
## 5                     GPT-4.1  201.20476       480  0.4191766
## 6            GPT-4.1-GPT-Image  190.82738       480  0.3975570
## 7                   ChatGPT-4o  206.76786       480  0.4307664
## 8                      o4-mini  255.87262       480  0.5330680
## 9                    Gemini-2.5  219.94881       480  0.4582267
## 10            Gemini-2.0-Flash  189.38214       480  0.3945461
## 11 Gemini-2.0-Flash-GPT-Image   77.80714       240  0.3241964
## 12                    Sonnet-4  201.98810       480  0.4208085
## 13                    Opus-4.1  118.85238       240  0.4952183
## 14                       GPT-5  308.62262       480  0.6429638
```

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

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

```r
print(collapsed_data)
```

```
##                         model      score max_score proportion
## 1                      Humans 4051.46667      7490  0.5409168
## 2                          o3  581.89048       900  0.6465450
## 3                o3-GPT-Image  666.93214      1200  0.5557768
## 4                      o3-Pro  597.87143       900  0.6643016
## 5                     GPT-4.1  258.61190       600  0.4310198
## 6           GPT-4.1-GPT-Image  233.07738       600  0.3884623
## 7                  ChatGPT-4o  259.99881       600  0.4333313
## 8                     o4-mini  319.88095       600  0.5331349
## 9                   Gemini-2.5  282.82381       600  0.4713730
## 10           Gemini-2.0-Flash  230.98214       600  0.3849702
## 11 Gemini-2.0-Flash-GPT-Image   96.84524       300  0.3228175
## 12                    Sonnet-4  258.39048       600  0.4306508
## 13                    Opus-4.1  163.81905       300  0.5460635
## 14                       GPT-5  400.82262       600  0.6680377
```

```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 proportion
## 1               Humans 952.09643      1525  0.6243255
## 2             o3-High 109.15000       180  0.6063889
## 3            o3-Medium  34.66667        60  0.5777778
## 4               o3-Low  37.63333        60  0.6272222
## 5           GPT-5-High  92.20000       120  0.7683333
## 6         GPT-5-Medium  38.00833        60  0.6334722
## 7            GPT-5-Low  33.60833        60  0.5601389
```

```
## 8        GPT-5-Minimal  22.18452        60  0.3697421
## 9         o4-mini-High  64.00833       120  0.5334028
## 10      o4-mini-Medium  55.27500       120  0.4606250
## 11   o3-GPT-Image-High 136.23333       240  0.5676389
## 12 o3-GPT-Image-Medium  29.83810        60  0.4973016
```

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

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

```
print(novel_reasoning_data)
```

```
##                  model     score max_score proportion
## 1               Humans 3099.3702      5965  0.5195927
## 2              o3-High  472.7405       720  0.6565840
## 3            o3-Medium  136.8440       240  0.5701835
## 4               o3-Low  126.6619       240  0.5277579
## 5           GPT-5-High  308.6226       480  0.6429638
## 6         GPT-5-Medium  140.1417       240  0.5839236
## 7            GPT-5-Low  119.2940       240  0.4970585
## 8        GPT-5-Minimal  100.2702       240  0.4177927
## 9         o4-mini-High  255.8726       480  0.5330680
## 10      o4-mini-Medium  237.9310       480  0.4956895
## 11   o3-GPT-Image-High  530.6988       960  0.5528113
## 12 o3-GPT-Image-Medium  136.7131       240  0.5696379
```

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

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

```
print(collapsed_reasoning_data)
```

```
##                  model     score max_score proportion
## 1               Humans 4051.4667      7490  0.5409168
## 2              o3-High  581.8905       900  0.6465450
## 3            o3-Medium  171.5107       300  0.5717024
## 4               o3-Low  164.2952       300  0.5476508
## 5           GPT-5-High  400.8226       600  0.6680377
## 6         GPT-5-Medium  178.1500       300  0.5938333
## 7            GPT-5-Low  152.9024       300  0.5096746
## 8        GPT-5-Minimal  122.4548       300  0.4081825
## 9         o4-mini-High  319.8810       600  0.5331349
## 10      o4-mini-Medium  293.2060       600  0.4886766
## 11   o3-GPT-Image-High  666.9321      1200  0.5557768
## 12 o3-GPT-Image-Medium  166.5512       300  0.5551706
```

Proportion Testing Function

```
# 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: o3 Single Context vs Multiple Context

##

```
##
## Comparison: o3 Family Single Context vs Multiple Context
## ================================================
## o3 Single Context: 191.7798/300 (0.639)
## o3 Multiple Context: 373.6155/600 (0.623)
## Difference:  0.017
## Chi-squared:  0.17
## P-value:  0.6805
## 95% CI: [ -0.053 ,  0.086 ]
## Significant:  NO
##
##
## Detailed Comparison: o3 Single Context vs Multiple Context
## ----------------------------------------
## Proportions:  0.639  vs  0.623
## Difference:  0.017
## Chi-squared:  0.17
## Degrees of freedom:  1
## P-value:  0.6805
## 95% CI: [ -0.053 ,  0.086 ]
## Significant:  NO
##
##
## Summary Table - o3 Single vs Multiple Context:
##
##
## comparison                              diff   p_value  significant
## ------------------------------------  ------  --------  ------------
## o3 Single Context vs Multiple Context   0.017    0.6805  FALSE
```

Comparison: o3 Pro Single Context vs Multiple Context

```
##
##
## Comparison: o3 Pro Family Single Context vs Multiple Context
## ================================================
## o3 Pro Single Context: 199.481/300 (0.665)
## o3 Pro Multiple Context: 396.3619/600 (0.661)
## Difference:  0.004
## Chi-squared:  0.003
## P-value:  0.9563
```

```
## 95% CI: [ -0.064 ,  0.072 ]
## Significant:  NO
##
##
## Detailed Comparison: o3 Pro Single Context vs Multiple Context
## ----------------------------------------
## Proportions:  0.665  vs  0.661
## Difference:  0.004
## Chi-squared:  0.003
## Degrees of freedom:  1
## P-value:  0.9563
## 95% CI: [ -0.064 ,  0.072 ]
## Significant:  NO
##
##
## Summary Table - o3 Pro Single vs Multiple Context:
##
##
## comparison                                  diff   p_value  significant
## ----------------------------------------  ------  --------  ------------
## o3 Pro Single Context vs Multiple Context   0.004    0.9563  FALSE
```

Comparison: Other OpenAI Single Context vs Multiple Context

Comparison: Gemini Single Context vs. Multiple Context

Comparison: Total Single Context vs Multiple Context

```
##
##
## Comparison: Total Single Context vs Multiple Context
## ================================================
## Total Single Context: 1311.76/2700 (0.486)
## Total Multiple Context: 1693.244/3300 (0.513)
## Difference:  -0.027
## Chi-squared:  4.308
## P-value:  0.03793
## 95% CI: [ -0.053 ,  -0.002 ]
## Significant:  YES (p < 0.05)
##
##
## Detailed Comparison: Total Single Context vs Multiple Context
## ----------------------------------------
```

```
## Proportions:  0.486  vs  0.513
## Difference:  -0.027
## Chi-squared:  4.308
## Degrees of freedom:  1
## P-value:  0.03793
## 95% CI: [ -0.053 ,  -0.002 ]
## Significant:  YES (p < 0.05)
##
##
## Summary Table - Total Single vs Multiple Context:
##
##
## comparison                                  diff   p_value  significant
## ----------------------------------------  -------  --------  ------------
## Total Single Context vs Multiple Context    -0.027    0.0379  TRUE
```

Comparison: Current Human Finke vs Original Finke

```
##
##
## Comparison: Current Human Finke vs Original Finke (Collapsed Exp 2 + Exp 3)
## ========================================================================
## Current Human Finke: 952.0964/1525 (0.624)
## Original Finke: 404/720 (0.561)
## Difference:  0.063
## Chi-squared:  7.909
## P-value:  0.004918
## 95% CI: [ 0.019 ,  0.108 ]
## Significant:  YES (p < 0.05)
##
##
## Detailed Comparison: Current Humans vs Original Finke
## --------------------------------------
## Proportions:  0.624  vs  0.561
## Difference:  0.063
## Chi-squared:  7.909
## Degrees of freedom:  1
## P-value:  0.004918
## 95% CI: [ 0.019 ,  0.108 ]
## Significant:  YES (p < 0.05)
```

```
##
##
## Summary Table - Human vs Original Finke:
##
##
## comparison                          diff   p_value  significant
## ---------------------------------  ------  --------  ------------
## Current Humans vs Original Finke    0.063    0.0049  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: 3099.37/5965 (0.52)
## Original Finke: 404/720 (0.561)
## Difference:  -0.042
## Chi-squared:  4.275
## P-value:  0.03867
## 95% CI: [ -0.081 ,  -0.002 ]
## Significant:  YES (p < 0.05)
##
##
## Detailed Comparison: Current Humans vs Original Finke
## ---------------------------------------
## Proportions:  0.52  vs  0.561
## Difference:  -0.042
## Chi-squared:  4.275
## Degrees of freedom:  1
## P-value:  0.03867
## 95% CI: [ -0.081 ,  -0.002 ]
## Significant:  YES (p < 0.05)
##
##
## Summary Table - Human vs Original Finke:
##
##
## comparison                            diff   p_value  significant
## ---------------------------------  -------  --------  ------------
## Current Humans vs Original Finke    -0.042    0.0387  TRUE
```

Comparison: Current Humans (collapsed) vs Original Finke

```
##
##
## Comparison: Current Human 48-Item Task vs Original Finke (Collapsed Exp 2 + Exp 3)
## ======================================================================
## Current Human Finke: 4051.467/7490 (0.541)
## Original Finke: 404/720 (0.561)
## Difference:  -0.02
## Chi-squared:  1
## P-value:  0.3174
## 95% CI: [ -0.059 ,  0.019 ]
## Significant:  NO
##
##
## Detailed Comparison: Current Humans vs Original Finke
## ----------------------------------------
## Proportions:  0.541  vs  0.561
## Difference:  -0.02
## Chi-squared:  1
## Degrees of freedom:  1
## P-value:  0.3174
## 95% CI: [ -0.059 ,  0.019 ]
## Significant:  NO
##
##
## Summary Table - Current Human (Collapsed) vs Original Finke:
##
##
## comparison                                    diff   p_value  significant
## --------------------------------------------  ------  --------  ------------
## Current Humans (collapsed) vs Original Finke    -0.02    0.3174  FALSE
```

Finke et al. Tasks - All Pairwise Comparisons

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

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

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

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

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

```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:  0.624  vs  0.606
## Difference:  0.018
## Chi-squared:  0.151
## Degrees of freedom:  1
## P-value:  0.6979
## 95% CI: [ -0.061 ,  0.096 ]
## Significant:  NO
##
##  Humans vs o3-GPT-Image
## ----------------------------------------
## Proportions:  0.624  vs  0.568
## Difference:  0.057
## Chi-squared:  2.584
## Degrees of freedom:  1
## P-value:  0.1079
## 95% CI: [ -0.013 ,  0.126 ]
## Significant:  NO
##
##  Humans vs o3-Pro
## ----------------------------------------
## Proportions:  0.624  vs  0.781
## Difference:  -0.157
## Chi-squared:  16.591
## Degrees of freedom:  1
## P-value:  0.00004636
## 95% CI: [ -0.225 ,  -0.089 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-4.1
## ----------------------------------------
## Proportions:  0.624  vs  0.478
## Difference:  0.146
## Chi-squared:  9.387
## Degrees of freedom:  1
## P-value:  0.002185
## 95% CI: [ 0.049 ,  0.243 ]
## Significant:  YES (p < 0.05)
```

```
##
##  Humans vs GPT-4.1-GPT-Image
## ---------------------------------------
## Proportions:  0.624  vs  0.352
## Difference:  0.272
## Chi-squared:  33.357
## Degrees of freedom:  1
## P-value:  0.000000007672
## 95% CI: [ 0.179 ,  0.366 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs ChatGPT-4o
## ---------------------------------------
## Proportions:  0.624  vs  0.444
## Difference:  0.181
## Chi-squared:  14.54
## Degrees of freedom:  1
## P-value:  0.0001372
## 95% CI: [ 0.084 ,  0.277 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o4-mini
## ---------------------------------------
## Proportions:  0.624  vs  0.533
## Difference:  0.091
## Chi-squared:  3.519
## Degrees of freedom:  1
## P-value:  0.06067
## 95% CI: [ -0.006 ,  0.188 ]
## Significant:  NO
##
##  Humans vs Gemini-2.5
## ---------------------------------------
## Proportions:  0.624  vs  0.524
## Difference:  0.1
## Chi-squared:  4.327
## Degrees of freedom:  1
## P-value:  0.03751
## 95% CI: [ 0.003 ,  0.197 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash
## ---------------------------------------
## Proportions:  0.624  vs  0.347
## Difference:  0.278
## Chi-squared:  34.708
## Degrees of freedom:  1
## P-value:  0.000000003831
## 95% CI: [ 0.185 ,  0.371 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.624  vs  0.317
```

```
## Difference:  0.307
## Chi-squared:  21.656
## Degrees of freedom:  1
## P-value:  0.000003261
## 95% CI: [ 0.178 ,  0.436 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Sonnet-4
## ----------------------------------------
## Proportions:  0.624  vs  0.47
## Difference:  0.154
## Chi-squared:  10.525
## Degrees of freedom:  1
## P-value:  0.001178
## 95% CI: [ 0.057 ,  0.251 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Opus-4.1
## ----------------------------------------
## Proportions:  0.624  vs  0.749
## Difference:  -0.125
## Chi-squared:  3.355
## Degrees of freedom:  1
## P-value:  0.06699
## 95% CI: [ -0.246 ,  -0.004 ]
## Significant:  NO
##
##  Humans vs GPT-5
## ----------------------------------------
## Proportions:  0.624  vs  0.768
## Difference:  -0.144
## Chi-squared:  9.34
## Degrees of freedom:  1
## P-value:  0.002242
## 95% CI: [ -0.228 ,  -0.06 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o3-GPT-Image
## ----------------------------------------
## Proportions:  0.606  vs  0.568
## Difference:  0.039
## Chi-squared:  0.486
## Degrees of freedom:  1
## P-value:  0.4856
## 95% CI: [ -0.061 ,  0.139 ]
## Significant:  NO
##
##  o3 vs o3-Pro
## ----------------------------------------
## Proportions:  0.606  vs  0.781
## Difference:  -0.175
## Chi-squared:  12.173
## Degrees of freedom:  1
## P-value:  0.000485
```

```
## 95% CI: [ -0.274 ,  -0.076 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs GPT-4.1
## ---------------------------------------
## Proportions:  0.606  vs  0.478
## Difference:  0.128
## Chi-squared:  4.272
## Degrees of freedom:  1
## P-value:  0.03874
## 95% CI: [ 0.007 ,  0.249 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs GPT-4.1-GPT-Image
## ---------------------------------------
## Proportions:  0.606  vs  0.352
## Difference:  0.254
## Chi-squared:  17.624
## Degrees of freedom:  1
## P-value:  0.00002692
## 95% CI: [ 0.136 ,  0.373 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs ChatGPT-4o
## ---------------------------------------
## Proportions:  0.606  vs  0.444
## Difference:  0.163
## Chi-squared:  7.044
## Degrees of freedom:  1
## P-value:  0.007955
## 95% CI: [ 0.042 ,  0.284 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o4-mini
## ---------------------------------------
## Proportions:  0.606  vs  0.533
## Difference:  0.073
## Chi-squared:  1.287
## Degrees of freedom:  1
## P-value:  0.2566
## 95% CI: [ -0.048 ,  0.194 ]
## Significant:  NO
##
##  o3 vs Gemini-2.5
## ---------------------------------------
## Proportions:  0.606  vs  0.524
## Difference:  0.082
## Chi-squared:  1.677
## Degrees of freedom:  1
## P-value:  0.1953
## 95% CI: [ -0.039 ,  0.204 ]
## Significant:  NO
##
##  o3 vs Gemini-2.0-Flash
```

```
## ----------------------------------------
## Proportions:  0.606  vs  0.347
## Difference:  0.26
## Chi-squared:  18.403
## Degrees of freedom:  1
## P-value:  0.00001788
## 95% CI: [ 0.142 ,  0.378 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.606  vs  0.317
## Difference:  0.289
## Chi-squared:  13.974
## Degrees of freedom:  1
## P-value:  0.0001854
## 95% CI: [ 0.14 ,  0.438 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Sonnet-4
## ----------------------------------------
## Proportions:  0.606  vs  0.47
## Difference:  0.136
## Chi-squared:  4.877
## Degrees of freedom:  1
## P-value:  0.02722
## 95% CI: [ 0.015 ,  0.258 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Opus-4.1
## ----------------------------------------
## Proportions:  0.606  vs  0.749
## Difference:  -0.143
## Chi-squared:  3.409
## Degrees of freedom:  1
## P-value:  0.06483
## 95% CI: [ -0.285 ,  -0.001 ]
## Significant:  NO
##
##  o3 vs GPT-5
## ----------------------------------------
## Proportions:  0.606  vs  0.768
## Difference:  -0.162
## Chi-squared:  7.838
## Degrees of freedom:  1
## P-value:  0.005117
## 95% CI: [ -0.273 ,  -0.051 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs o3-Pro
## ----------------------------------------
## Proportions:  0.568  vs  0.781
## Difference:  -0.214
## Chi-squared:  19.989
```

```
## Degrees of freedom:  1
## P-value:  0.000007791
## 95% CI: [ -0.306 ,  -0.122 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs GPT-4.1
## ----------------------------------------
## Proportions:  0.568  vs  0.478
## Difference:  0.089
## Chi-squared:  2.217
## Degrees of freedom:  1
## P-value:  0.1365
## 95% CI: [ -0.026 ,  0.205 ]
## Significant:  NO
##
##  o3-GPT-Image vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.568  vs  0.352
## Difference:  0.216
## Chi-squared:  14.02
## Degrees of freedom:  1
## P-value:  0.0001809
## 95% CI: [ 0.103 ,  0.328 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.568  vs  0.444
## Difference:  0.124
## Chi-squared:  4.453
## Degrees of freedom:  1
## P-value:  0.03485
## 95% CI: [ 0.009 ,  0.239 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs o4-mini
## ----------------------------------------
## Proportions:  0.568  vs  0.533
## Difference:  0.034
## Chi-squared:  0.254
## Degrees of freedom:  1
## P-value:  0.6144
## 95% CI: [ -0.081 ,  0.15 ]
## Significant:  NO
##
##  o3-GPT-Image vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.568  vs  0.524
## Difference:  0.044
## Chi-squared:  0.453
## Degrees of freedom:  1
## P-value:  0.5007
## 95% CI: [ -0.072 ,  0.159 ]
## Significant:  NO
```

```
##
##  o3-GPT-Image vs Gemini-2.0-Flash
## ---------------------------------------
## Proportions:  0.568  vs  0.347
## Difference:  0.221
## Chi-squared:  14.756
## Degrees of freedom:  1
## P-value:  0.0001224
## 95% CI: [ 0.109 ,  0.333 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.568  vs  0.317
## Difference:  0.25
## Chi-squared:  11.066
## Degrees of freedom:  1
## P-value:  0.0008795
## 95% CI: [ 0.107 ,  0.394 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Sonnet-4
## ---------------------------------------
## Proportions:  0.568  vs  0.47
## Difference:  0.098
## Chi-squared:  2.685
## Degrees of freedom:  1
## P-value:  0.1013
## 95% CI: [ -0.018 ,  0.213 ]
## Significant:  NO
##
##  o3-GPT-Image vs Opus-4.1
## ---------------------------------------
## Proportions:  0.568  vs  0.749
## Difference:  -0.182
## Chi-squared:  5.895
## Degrees of freedom:  1
## P-value:  0.01519
## 95% CI: [ -0.319 ,  -0.045 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs GPT-5
## ---------------------------------------
## Proportions:  0.568  vs  0.768
## Difference:  -0.201
## Chi-squared:  13.043
## Degrees of freedom:  1
## P-value:  0.0003044
## 95% CI: [ -0.305 ,  -0.096 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-4.1
## ---------------------------------------
## Proportions:  0.781  vs  0.478
```

```
## Difference:  0.303
## Chi-squared:  28.139
## Degrees of freedom:  1
## P-value:  0.0000001129
## 95% CI: [ 0.188 ,  0.418 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.781  vs  0.352
## Difference:  0.429
## Chi-squared:  53.986
## Degrees of freedom:  1
## P-value:  0.000000000000202
## 95% CI: [ 0.318 ,  0.541 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.781  vs  0.444
## Difference:  0.338
## Chi-squared:  34.487
## Degrees of freedom:  1
## P-value:  0.000000004291
## 95% CI: [ 0.223 ,  0.452 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o4-mini
## ----------------------------------------
## Proportions:  0.781  vs  0.533
## Difference:  0.248
## Chi-squared:  19.303
## Degrees of freedom:  1
## P-value:  0.00001115
## 95% CI: [ 0.133 ,  0.363 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.781  vs  0.524
## Difference:  0.257
## Chi-squared:  20.715
## Degrees of freedom:  1
## P-value:  0.000005329
## 95% CI: [ 0.143 ,  0.372 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.781  vs  0.347
## Difference:  0.435
## Chi-squared:  55.27
## Degrees of freedom:  1
## P-value:  0.0000000000001051
```

```
## 95% CI: [ 0.323 ,  0.546 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.781  vs  0.317
## Difference:  0.464
## Chi-squared:  41.481
## Degrees of freedom:  1
## P-value:  0.000000000119
## 95% CI: [ 0.321 ,  0.608 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Sonnet-4
## ---------------------------------------
## Proportions:  0.781  vs  0.47
## Difference:  0.311
## Chi-squared:  29.613
## Degrees of freedom:  1
## P-value:  0.00000005274
## 95% CI: [ 0.197 ,  0.426 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Opus-4.1
## ---------------------------------------
## Proportions:  0.781  vs  0.749
## Difference:  0.032
## Chi-squared:  0.112
## Degrees of freedom:  1
## P-value:  0.7379
## 95% CI: [ -0.104 ,  0.168 ]
## Significant:  NO
##
##  o3-Pro vs GPT-5
## ---------------------------------------
## Proportions:  0.781  vs  0.768
## Difference:  0.013
## Chi-squared:  0.016
## Degrees of freedom:  1
## P-value:  0.9002
## 95% CI: [ -0.091 ,  0.117 ]
## Significant:  NO
##
##  GPT-4.1 vs GPT-4.1-GPT-Image
## ---------------------------------------
## Proportions:  0.478  vs  0.352
## Difference:  0.126
## Chi-squared:  3.439
## Degrees of freedom:  1
## P-value:  0.06366
## 95% CI: [ -0.006 ,  0.258 ]
## Significant:  NO
##
##  GPT-4.1 vs ChatGPT-4o
```

```
## ----------------------------------------
## Proportions:  0.478  vs  0.444
## Difference:  0.035
## Chi-squared:  0.169
## Degrees of freedom:  1
## P-value:  0.6809
## 95% CI: [ -0.1 ,  0.169 ]
## Significant:  NO
##
##  GPT-4.1 vs o4-mini
## ----------------------------------------
## Proportions:  0.478  vs  0.533
## 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
## ----------------------------------------
## Proportions:  0.478  vs  0.524
## Difference:  -0.046
## Chi-squared:  0.333
## Degrees of freedom:  1
## P-value:  0.5641
## 95% CI: [ -0.18 ,  0.089 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.478  vs  0.347
## Difference:  0.132
## Chi-squared:  3.77
## Degrees of freedom:  1
## P-value:  0.05219
## 95% CI: [ 0 ,  0.264 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.478  vs  0.317
## Difference:  0.161
## Chi-squared:  3.615
## Degrees of freedom:  1
## P-value:  0.05727
## 95% CI: [ 0.001 ,  0.321 ]
## Significant:  NO
##
##  GPT-4.1 vs Sonnet-4
## ----------------------------------------
## Proportions:  0.478  vs  0.47
## Difference:  0.008
## Chi-squared:  0
```

```
## Degrees of freedom:  1
## P-value:  0.9995
## 95% CI: [ -0.126 ,  0.143 ]
## Significant:  NO
##
##  GPT-4.1 vs Opus-4.1
## ----------------------------------------
## Proportions:  0.478  vs  0.749
## Difference:  -0.271
## Chi-squared:  10.902
## Degrees of freedom:  1
## P-value:  0.0009607
## 95% CI: [ -0.425 ,  -0.117 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs GPT-5
## ----------------------------------------
## Proportions:  0.478  vs  0.768
## Difference:  -0.29
## Chi-squared:  20.266
## Degrees of freedom:  1
## P-value:  0.000006738
## 95% CI: [ -0.415 ,  -0.165 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.352  vs  0.444
## Difference:  -0.092
## Chi-squared:  1.733
## Degrees of freedom:  1
## P-value:  0.1881
## 95% CI: [ -0.223 ,  0.04 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs o4-mini
## ----------------------------------------
## Proportions:  0.352  vs  0.533
## Difference:  -0.181
## Chi-squared:  7.277
## Degrees of freedom:  1
## P-value:  0.006983
## 95% CI: [ -0.313 ,  -0.049 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.352  vs  0.524
## Difference:  -0.172
## Chi-squared:  6.519
## Degrees of freedom:  1
## P-value:  0.01067
## 95% CI: [ -0.304 ,  -0.04 ]
## Significant:  YES (p < 0.05)
```

```
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash
## ---------------------------------------
## Proportions:  0.352  vs  0.347
## Difference:  0.005
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.121 ,  0.131 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.352  vs  0.317
## Difference:  0.035
## Chi-squared:  0.088
## Degrees of freedom:  1
## P-value:  0.7662
## 95% CI: [ -0.123 ,  0.193 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Sonnet-4
## ---------------------------------------
## Proportions:  0.352  vs  0.47
## Difference:  -0.118
## Chi-squared:  2.977
## Degrees of freedom:  1
## P-value:  0.08444
## 95% CI: [ -0.25 ,  0.014 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Opus-4.1
## ---------------------------------------
## Proportions:  0.352  vs  0.749
## Difference:  -0.397
## Chi-squared:  23.722
## Degrees of freedom:  1
## P-value:  0.000001113
## 95% CI: [ -0.549 ,  -0.246 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs GPT-5
## ---------------------------------------
## Proportions:  0.352  vs  0.768
## Difference:  -0.416
## Chi-squared:  40.523
## Degrees of freedom:  1
## P-value:  0.0000000001944
## 95% CI: [ -0.539 ,  -0.294 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs o4-mini
## ---------------------------------------
## Proportions:  0.444  vs  0.533
```

```
## Difference:  -0.09
## Chi-squared:  1.594
## Degrees of freedom:  1
## P-value:  0.2067
## 95% CI: [ -0.224 ,  0.044 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.444  vs  0.524
## Difference:  -0.08
## Chi-squared:  1.247
## Degrees of freedom:  1
## P-value:  0.2642
## 95% CI: [ -0.215 ,  0.054 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.444  vs  0.347
## Difference:  0.097
## Chi-squared:  1.97
## Degrees of freedom:  1
## P-value:  0.1604
## 95% CI: [ -0.035 ,  0.228 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.444  vs  0.317
## Difference:  0.126
## Chi-squared:  2.155
## Degrees of freedom:  1
## P-value:  0.1421
## 95% CI: [ -0.034 ,  0.286 ]
## Significant:  NO
##
##  ChatGPT-4o vs Sonnet-4
## ----------------------------------------
## Proportions:  0.444  vs  0.47
## Difference:  -0.026
## Chi-squared:  0.079
## Degrees of freedom:  1
## P-value:  0.7784
## 95% CI: [ -0.161 ,  0.108 ]
## Significant:  NO
##
##  ChatGPT-4o vs Opus-4.1
## ----------------------------------------
## Proportions:  0.444  vs  0.749
## Difference:  -0.306
## Chi-squared:  13.884
## Degrees of freedom:  1
## P-value:  0.0001944
```

```
## 95% CI: [ -0.46 ,  -0.152 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs GPT-5
## ---------------------------------------
## Proportions:  0.444  vs  0.768
## Difference:  -0.325
## Chi-squared:  25.157
## Degrees of freedom:  1
## P-value:  0.0000005284
## 95% CI: [ -0.45 ,  -0.2 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Gemini-2.5
## ---------------------------------------
## Proportions:  0.533  vs  0.524
## Difference:  0.009
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  0.9862
## 95% CI: [ -0.125 ,  0.144 ]
## Significant:  NO
##
##  o4-mini vs Gemini-2.0-Flash
## ---------------------------------------
## Proportions:  0.533  vs  0.347
## Difference:  0.187
## Chi-squared:  7.75
## Degrees of freedom:  1
## P-value:  0.005371
## 95% CI: [ 0.055 ,  0.318 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.533  vs  0.317
## Difference:  0.216
## Chi-squared:  6.672
## Degrees of freedom:  1
## P-value:  0.009792
## 95% CI: [ 0.056 ,  0.376 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Sonnet-4
## ---------------------------------------
## Proportions:  0.533  vs  0.47
## Difference:  0.063
## Chi-squared:  0.727
## Degrees of freedom:  1
## P-value:  0.3938
## 95% CI: [ -0.071 ,  0.198 ]
## Significant:  NO
##
##  o4-mini vs Opus-4.1
```

```
## ----------------------------------------
## Proportions:  0.533  vs  0.749
## Difference:  -0.216
## Chi-squared:  6.937
## Degrees of freedom:  1
## P-value:  0.008443
## 95% CI: [ -0.37 ,  -0.062 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs GPT-5
## ----------------------------------------
## Proportions:  0.533  vs  0.768
## Difference:  -0.235
## Chi-squared:  13.557
## Degrees of freedom:  1
## P-value:  0.0002314
## 95% CI: [ -0.36 ,  -0.11 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.524  vs  0.347
## Difference:  0.177
## Chi-squared:  6.968
## Degrees of freedom:  1
## P-value:  0.008299
## 95% CI: [ 0.046 ,  0.309 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5 vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.524  vs  0.317
## Difference:  0.207
## Chi-squared:  6.081
## Degrees of freedom:  1
## P-value:  0.01367
## 95% CI: [ 0.046 ,  0.367 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5 vs Sonnet-4
## ----------------------------------------
## Proportions:  0.524  vs  0.47
## Difference:  0.054
## Chi-squared:  0.499
## Degrees of freedom:  1
## P-value:  0.4799
## 95% CI: [ -0.081 ,  0.189 ]
## Significant:  NO
##
##  Gemini-2.5 vs Opus-4.1
## ----------------------------------------
## Proportions:  0.524  vs  0.749
## Difference:  -0.225
## Chi-squared:  7.555
```

```
## Degrees of freedom:  1
## P-value:  0.005984
## 95% CI: [ -0.379 ,  -0.072 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5 vs GPT-5
## ----------------------------------------
## Proportions:  0.524  vs  0.768
## Difference:  -0.244
## Chi-squared:  14.621
## Degrees of freedom:  1
## P-value:  0.0001314
## 95% CI: [ -0.37 ,  -0.119 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.347  vs  0.317
## Difference:  0.029
## Chi-squared:  0.051
## Degrees of freedom:  1
## P-value:  0.8215
## 95% CI: [ -0.128 ,  0.187 ]
## Significant:  NO
##
##  Gemini-2.0-Flash vs Sonnet-4
## ----------------------------------------
## Proportions:  0.347  vs  0.47
## Difference:  -0.123
## Chi-squared:  3.286
## Degrees of freedom:  1
## P-value:  0.06989
## 95% CI: [ -0.255 ,  0.008 ]
## Significant:  NO
##
##  Gemini-2.0-Flash vs Opus-4.1
## ----------------------------------------
## Proportions:  0.347  vs  0.749
## Difference:  -0.403
## Chi-squared:  24.406
## Degrees of freedom:  1
## P-value:  0.0000007802
## 95% CI: [ -0.554 ,  -0.251 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs GPT-5
## ----------------------------------------
## Proportions:  0.347  vs  0.768
## Difference:  -0.422
## Chi-squared:  41.552
## Degrees of freedom:  1
## P-value:  0.0000000001148
## 95% CI: [ -0.544 ,  -0.3 ]
## Significant:  YES (p < 0.05)
```

```
##
##  Gemini-2.0-Flash-GPT-Image vs Sonnet-4
## ---------------------------------------
## Proportions:  0.317  vs  0.47
## Difference:  -0.153
## Chi-squared:  3.23
## Degrees of freedom:  1
## P-value:  0.07229
## 95% CI: [ -0.313 ,  0.008 ]
## Significant:  NO
##
##  Gemini-2.0-Flash-GPT-Image vs Opus-4.1
## ---------------------------------------
## Proportions:  0.317  vs  0.749
## Difference:  -0.432
## Chi-squared:  20.807
## Degrees of freedom:  1
## P-value:  0.000005079
## 95% CI: [ -0.61 ,  -0.255 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-GPT-Image vs GPT-5
## ---------------------------------------
## Proportions:  0.317  vs  0.768
## Difference:  -0.451
## Chi-squared:  32.584
## Degrees of freedom:  1
## P-value:  0.00000001141
## 95% CI: [ -0.603 ,  -0.299 ]
## Significant:  YES (p < 0.05)
##
##  Sonnet-4 vs Opus-4.1
## ---------------------------------------
## Proportions:  0.47  vs  0.749
## Difference:  -0.279
## Chi-squared:  11.585
## Degrees of freedom:  1
## P-value:  0.000665
## 95% CI: [ -0.433 ,  -0.126 ]
## Significant:  YES (p < 0.05)
##
##  Sonnet-4 vs GPT-5
## ---------------------------------------
## Proportions:  0.47  vs  0.768
## Difference:  -0.298
## Chi-squared:  21.397
## Degrees of freedom:  1
## P-value:  0.000003734
## 95% CI: [ -0.424 ,  -0.173 ]
## Significant:  YES (p < 0.05)
##
##  Opus-4.1 vs GPT-5
## ---------------------------------------
## Proportions:  0.749  vs  0.768
```

```
## Difference:  -0.019
## Chi-squared:  0.009
## Degrees of freedom:  1
## P-value:  0.9244
## 95% CI: [ -0.165 ,  0.127 ]
## Significant:  NO
```

```
# 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:
```

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

```
##
##
##              comparison                                      diff   chi_squared   p_value  sig
## ------------  ----------------------------------------------  -------  ------------  --------  ---
## X-squared     Humans vs o3                                     0.018     0.1506854    0.6979  FALS
## X-squared1    Humans vs o3-GPT-Image                           0.057     2.5840139    0.1079  FALS
## X-squared2    Humans vs o3-Pro                                -0.157    16.5913863    0.0000  TRU
## X-squared3    Humans vs GPT-4.1                                0.146     9.3870484    0.0022  TRU
## X-squared4    Humans vs GPT-4.1-GPT-Image                      0.272    33.3565263    0.0000  TRU
## X-squared5    Humans vs ChatGPT-4o                             0.181    14.5398296    0.0001  TRU
## X-squared6    Humans vs o4-mini                                0.091     3.5189352    0.0607  FALS
## X-squared7    Humans vs Gemini-2.5                             0.100     4.3270555    0.0375  TRU
## X-squared8    Humans vs Gemini-2.0-Flash                       0.278    34.7078541    0.0000  TRU
## X-squared9    Humans vs Gemini-2.0-Flash-GPT-Image             0.307    21.6564329    0.0000  TRU
## X-squared10   Humans vs Sonnet-4                               0.154    10.5252073    0.0012  TRU
## X-squared11   Humans vs Opus-4.1                              -0.125     3.3553168    0.0670  FALS
## X-squared12   Humans vs GPT-5                                 -0.144     9.3403689    0.0022  TRU
## X-squared13   o3 vs o3-GPT-Image                              0.039     0.4863142    0.4856  FALS
## X-squared14   o3 vs o3-Pro                                   -0.175    12.1726124    0.0005  TRU
## X-squared15   o3 vs GPT-4.1                                   0.128     4.2722572    0.0387  TRU
## X-squared16   o3 vs GPT-4.1-GPT-Image                         0.254    17.6235408    0.0000  TRU
## X-squared17   o3 vs ChatGPT-4o                                0.163     7.0435675    0.0080  TRU
## X-squared18   o3 vs o4-mini                                   0.073     1.2867842    0.2566  FALS
## X-squared19   o3 vs Gemini-2.5                                0.082     1.6772292    0.1953  FALS
## X-squared20   o3 vs Gemini-2.0-Flash                          0.260    18.4026823    0.0000  TRU
## X-squared21   o3 vs Gemini-2.0-Flash-GPT-Image                0.289    13.9737978    0.0002  TRU
## X-squared22   o3 vs Sonnet-4                                  0.136     4.8766338    0.0272  TRU
## X-squared23   o3 vs Opus-4.1                                 -0.143     3.4092490    0.0648  FALS
## X-squared24   o3 vs GPT-5                                    -0.162     7.8377188    0.0051  TRU
## X-squared25   o3-GPT-Image vs o3-Pro                         -0.214    19.9885547    0.0000  TRU
## X-squared26   o3-GPT-Image vs GPT-4.1                         0.089     2.2170007    0.1365  FALS
## X-squared27   o3-GPT-Image vs GPT-4.1-GPT-Image               0.216    14.0198163    0.0002  TRU
## X-squared28   o3-GPT-Image vs ChatGPT-4o                      0.124     4.4527192    0.0348  TRU
## X-squared29   o3-GPT-Image vs o4-mini                         0.034     0.2538412    0.6144  FALS
```

```
## X-squared30   o3-GPT-Image vs Gemini-2.5                          0.044     0.4534449    0.5007  FALS
## X-squared31   o3-GPT-Image vs Gemini-2.0-Flash                    0.221    14.7559405    0.0001  TRUE
## X-squared32   o3-GPT-Image vs Gemini-2.0-Flash-GPT-Image         0.250    11.0655531    0.0009  TRUE
## X-squared33   o3-GPT-Image vs Sonnet-4                           0.098     2.6846866    0.1013  FALS
## X-squared34   o3-GPT-Image vs Opus-4.1                          -0.182     5.8948728    0.0152  TRUE
## X-squared35   o3-GPT-Image vs GPT-5                            -0.201    13.0430968    0.0003  TRUE
## X-squared36   o3-Pro vs GPT-4.1                                 0.303    28.1394453    0.0000  TRUE
## X-squared37   o3-Pro vs GPT-4.1-GPT-Image                       0.429    53.9856801    0.0000  TRUE
## X-squared38   o3-Pro vs ChatGPT-4o                              0.338    34.4868956    0.0000  TRUE
## X-squared39   o3-Pro vs o4-mini                                 0.248    19.3034158    0.0000  TRUE
## X-squared40   o3-Pro vs Gemini-2.5                              0.257    20.7152420    0.0000  TRUE
## X-squared41   o3-Pro vs Gemini-2.0-Flash                        0.435    55.2697971    0.0000  TRUE
## X-squared42   o3-Pro vs Gemini-2.0-Flash-GPT-Image             0.464    41.4805449    0.0000  TRUE
## X-squared43   o3-Pro vs Sonnet-4                               0.311    29.6132030    0.0000  TRUE
## X-squared44   o3-Pro vs Opus-4.1                               0.032     0.1119553    0.7379  FALS
## X-squared45   o3-Pro vs GPT-5                                  0.013     0.0157140    0.9002  FALS
## X-squared46   GPT-4.1 vs GPT-4.1-GPT-Image                     0.126     3.4392498    0.0637  FALS
## X-squared47   GPT-4.1 vs ChatGPT-4o                            0.035     0.1691661    0.6809  FALS
## X-squared48   GPT-4.1 vs o4-mini                              -0.055     0.5229617    0.4696  FALS
## X-squared49   GPT-4.1 vs Gemini-2.5                           -0.046     0.3326976    0.5641  FALS
## X-squared50   GPT-4.1 vs Gemini-2.0-Flash                      0.132     3.7695554    0.0522  FALS
## X-squared51   GPT-4.1 vs Gemini-2.0-Flash-GPT-Image           0.161     3.6146902    0.0573  FALS
## X-squared52   GPT-4.1 vs Sonnet-4                             0.008     0.0000004    0.9995  FALS
## X-squared53   GPT-4.1 vs Opus-4.1                            -0.271    10.9019022    0.0010  TRUE
## X-squared54   GPT-4.1 vs GPT-5                              -0.290    20.2663073    0.0000  TRUE
## X-squared55   GPT-4.1-GPT-Image vs ChatGPT-4o                -0.092     1.7326601    0.1881  FALS
## X-squared56   GPT-4.1-GPT-Image vs o4-mini                   -0.181     7.2772362    0.0070  TRUE
## X-squared57   GPT-4.1-GPT-Image vs Gemini-2.5                -0.172     6.5191820    0.0107  TRUE
## X-squared58   GPT-4.1-GPT-Image vs Gemini-2.0-Flash           0.005     0.0000000    1.0000  FALS
## X-squared59   GPT-4.1-GPT-Image vs Gemini-2.0-Flash-GPT-Image  0.035     0.0884368    0.7662  FALS
## X-squared60   GPT-4.1-GPT-Image vs Sonnet-4                   -0.118     2.9773091    0.0844  FALS
## X-squared61   GPT-4.1-GPT-Image vs Opus-4.1                  -0.397    23.7215796    0.0000  TRUE
## X-squared62   GPT-4.1-GPT-Image vs GPT-5                     -0.416    40.5226268    0.0000  TRUE
## X-squared63   ChatGPT-4o vs o4-mini                          -0.090     1.5941300    0.2067  FALS
## X-squared64   ChatGPT-4o vs Gemini-2.5                       -0.080     1.2466387    0.2642  FALS
## X-squared65   ChatGPT-4o vs Gemini-2.0-Flash                  0.097     1.9702959    0.1604  FALS
## X-squared66   ChatGPT-4o vs Gemini-2.0-Flash-GPT-Image       0.126     2.1553507    0.1421  FALS
## X-squared67   ChatGPT-4o vs Sonnet-4                         -0.026     0.0791759    0.7784  FALS
## X-squared68   ChatGPT-4o vs Opus-4.1                        -0.306    13.8841624    0.0002  TRUE
## X-squared69   ChatGPT-4o vs GPT-5                           -0.325    25.1573439    0.0000  TRUE
## X-squared70   o4-mini vs Gemini-2.5                          0.009     0.0002973    0.9862  FALS
## X-squared71   o4-mini vs Gemini-2.0-Flash                    0.187     7.7500844    0.0054  TRUE
## X-squared72   o4-mini vs Gemini-2.0-Flash-GPT-Image         0.216     6.6723813    0.0098  TRUE
## X-squared73   o4-mini vs Sonnet-4                           0.063     0.7273186    0.3938  FALS
## X-squared74   o4-mini vs Opus-4.1                          -0.216     6.9370294    0.0084  TRUE
## X-squared75   o4-mini vs GPT-5                             -0.235    13.5574456    0.0002  TRUE
## X-squared76   Gemini-2.5 vs Gemini-2.0-Flash                0.177     6.9678882    0.0083  TRUE
## X-squared77   Gemini-2.5 vs Gemini-2.0-Flash-GPT-Image     0.207     6.0805884    0.0137  TRUE
## X-squared78   Gemini-2.5 vs Sonnet-4                       0.054     0.4991774    0.4799  FALS
## X-squared79   Gemini-2.5 vs Opus-4.1                      -0.225     7.5550003    0.0060  TRUE
## X-squared80   Gemini-2.5 vs GPT-5                         -0.244    14.6208874    0.0001  TRUE
## X-squared81   Gemini-2.0-Flash vs Gemini-2.0-Flash-GPT-Image  0.029     0.0509296    0.8215  FALS
## X-squared82   Gemini-2.0-Flash vs Sonnet-4                 -0.123     3.2855008    0.0699  FALS
## X-squared83   Gemini-2.0-Flash vs Opus-4.1                -0.403    24.4061969    0.0000  TRUE
```

```
## X-squared84   Gemini-2.0-Flash vs GPT-5                          -0.422    41.5521944    0.0000  TRU
## X-squared85   Gemini-2.0-Flash-GPT-Image vs Sonnet-4             -0.153     3.2303241    0.0723  FALS
## X-squared86   Gemini-2.0-Flash-GPT-Image vs Opus-4.1             -0.432    20.8071523    0.0000  TRU
## X-squared87   Gemini-2.0-Flash-GPT-Image vs GPT-5                -0.451    32.5840929    0.0000  TRU
## X-squared88   Sonnet-4 vs Opus-4.1                               -0.279    11.5846605    0.0007  TRU
## X-squared89   Sonnet-4 vs GPT-5                                  -0.298    21.3968405    0.0000  TRU
## X-squared90   Opus-4.1 vs GPT-5                                  -0.019     0.0090038    0.9244  FALS
```

48 Novel Tasks - All Pairwise Comparisons

```
# 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:
```

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

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

```
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:  0.52  vs  0.657
## Difference:  -0.137
## Chi-squared:  47.906
## Degrees of freedom:  1
## P-value:  0.000000000004471
## 95% CI: [ -0.175 ,  -0.099 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o3-GPT-Image
## ----------------------------------------
## Proportions:  0.52  vs  0.553
## Difference:  -0.033
## Chi-squared:  3.527
## Degrees of freedom:  1
## P-value:  0.06039
## 95% CI: [ -0.068 ,  0.001 ]
```

```
## Significant:  NO
##
##  Humans vs o3-Pro
## ----------------------------------------
## Proportions:  0.52  vs  0.635
## Difference:  -0.115
## Chi-squared:  33.917
## Degrees of freedom:  1
## P-value:  0.000000005752
## 95% CI: [ -0.154 ,  -0.077 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-4.1
## ----------------------------------------
## Proportions:  0.52  vs  0.419
## Difference:  0.1
## Chi-squared:  17.529
## Degrees of freedom:  1
## P-value:  0.00002829
## 95% CI: [ 0.053 ,  0.147 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.52  vs  0.398
## Difference:  0.122
## Chi-squared:  25.99
## Degrees of freedom:  1
## P-value:  0.0000003432
## 95% CI: [ 0.075 ,  0.169 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.52  vs  0.431
## Difference:  0.089
## Chi-squared:  13.677
## Degrees of freedom:  1
## P-value:  0.0002171
## 95% CI: [ 0.042 ,  0.136 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o4-mini
## ----------------------------------------
## Proportions:  0.52  vs  0.533
## Difference:  -0.013
## Chi-squared:  0.271
## Degrees of freedom:  1
## P-value:  0.6023
## 95% CI: [ -0.061 ,  0.034 ]
## Significant:  NO
##
##  Humans vs Gemini-2.5
## ----------------------------------------
```

```
## Proportions:  0.52  vs  0.458
## Difference:  0.061
## Chi-squared:  6.454
## Degrees of freedom:  1
## P-value:  0.01107
## 95% CI: [ 0.014 ,  0.109 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash
## ---------------------------------------
## Proportions:  0.52  vs  0.395
## Difference:  0.125
## Chi-squared:  27.3
## Degrees of freedom:  1
## P-value:  0.0000001742
## 95% CI: [ 0.078 ,  0.172 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.52  vs  0.324
## Difference:  0.195
## Chi-squared:  34.478
## Degrees of freedom:  1
## P-value:  0.000000004312
## 95% CI: [ 0.133 ,  0.258 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Sonnet-4
## ---------------------------------------
## Proportions:  0.52  vs  0.421
## Difference:  0.099
## Chi-squared:  16.958
## Degrees of freedom:  1
## P-value:  0.00003822
## 95% CI: [ 0.052 ,  0.146 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Opus-4.1
## ---------------------------------------
## Proportions:  0.52  vs  0.495
## Difference:  0.024
## Chi-squared:  0.456
## Degrees of freedom:  1
## P-value:  0.4996
## 95% CI: [ -0.042 ,  0.091 ]
## Significant:  NO
##
##  Humans vs GPT-5
## ---------------------------------------
## Proportions:  0.52  vs  0.643
## Difference:  -0.123
## Chi-squared:  26.644
## Degrees of freedom:  1
```

```
## P-value:  0.0000002446
## 95% CI: [ -0.169 ,  -0.078 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o3-GPT-Image
## ----------------------------------------
## Proportions:  0.657  vs  0.553
## Difference:  0.104
## Chi-squared:  17.991
## Degrees of freedom:  1
## P-value:  0.0000222
## 95% CI: [ 0.056 ,  0.152 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o3-Pro
## ----------------------------------------
## Proportions:  0.657  vs  0.635
## Difference:  0.022
## Chi-squared:  0.641
## Degrees of freedom:  1
## P-value:  0.4235
## 95% CI: [ -0.029 ,  0.072 ]
## Significant:  NO
##
##  o3 vs GPT-4.1
## ----------------------------------------
## Proportions:  0.657  vs  0.419
## Difference:  0.237
## Chi-squared:  64.97
## Degrees of freedom:  1
## P-value:  0.0000000000000007605
## 95% CI: [ 0.18 ,  0.295 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.657  vs  0.398
## Difference:  0.259
## Chi-squared:  77.126
## Degrees of freedom:  1
## P-value:  0.00000000000000001604
## 95% CI: [ 0.201 ,  0.317 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.657  vs  0.431
## Difference:  0.226
## Chi-squared:  58.879
## Degrees of freedom:  1
## P-value:  0.00000000000001677
## 95% CI: [ 0.168 ,  0.284 ]
## Significant:  YES (p < 0.05)
##
```

```
##  o3 vs o4-mini
## ---------------------------------------
## Proportions:  0.657  vs  0.533
## Difference:  0.124
## Chi-squared:  17.907
## Degrees of freedom:  1
## P-value:  0.00002319
## 95% CI: [ 0.065 ,  0.182 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.5
## ---------------------------------------
## Proportions:  0.657  vs  0.458
## Difference:  0.198
## Chi-squared:  45.625
## Degrees of freedom:  1
## P-value:  0.00000000001432
## 95% CI: [ 0.14 ,  0.257 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.0-Flash
## ---------------------------------------
## Proportions:  0.657  vs  0.395
## Difference:  0.262
## Chi-squared:  78.902
## Degrees of freedom:  1
## P-value:  0.00000000000000006527
## 95% CI: [ 0.204 ,  0.32 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.657  vs  0.324
## Difference:  0.332
## Chi-squared:  79.95
## Degrees of freedom:  1
## P-value:  0.00000000000000003841
## 95% CI: [ 0.261 ,  0.404 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Sonnet-4
## ---------------------------------------
## Proportions:  0.657  vs  0.421
## Difference:  0.236
## Chi-squared:  64.094
## Degrees of freedom:  1
## P-value:  0.000000000000001186
## 95% CI: [ 0.178 ,  0.294 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Opus-4.1
## ---------------------------------------
## Proportions:  0.657  vs  0.495
## Difference:  0.161
```

```
## Chi-squared:  19.143
## Degrees of freedom:  1
## P-value:  0.00001213
## 95% CI: [ 0.086 ,  0.236 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs GPT-5
## ---------------------------------------
## Proportions:  0.657  vs  0.643
## Difference:  0.014
## Chi-squared:  0.179
## Degrees of freedom:  1
## P-value:  0.6722
## 95% CI: [ -0.043 ,  0.07 ]
## Significant:  NO
##
##  o3-GPT-Image vs o3-Pro
## ---------------------------------------
## Proportions:  0.553  vs  0.635
## Difference:  -0.082
## Chi-squared:  11.141
## Degrees of freedom:  1
## P-value:  0.0008445
## 95% CI: [ -0.131 ,  -0.034 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs GPT-4.1
## ---------------------------------------
## Proportions:  0.553  vs  0.419
## Difference:  0.134
## Chi-squared:  22.333
## Degrees of freedom:  1
## P-value:  0.000002292
## 95% CI: [ 0.078 ,  0.189 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs GPT-4.1-GPT-Image
## ---------------------------------------
## Proportions:  0.553  vs  0.398
## Difference:  0.155
## Chi-squared:  30.235
## Degrees of freedom:  1
## P-value:  0.00000003827
## 95% CI: [ 0.1 ,  0.211 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs ChatGPT-4o
## ---------------------------------------
## Proportions:  0.553  vs  0.431
## Difference:  0.122
## Chi-squared:  18.591
## Degrees of freedom:  1
## P-value:  0.00001619
## 95% CI: [ 0.066 ,  0.178 ]
```

```
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs o4-mini
## ---------------------------------------
## Proportions:  0.553  vs  0.533
## Difference:  0.02
## Chi-squared:  0.427
## Degrees of freedom:  1
## P-value:  0.5136
## 95% CI: [ -0.036 ,  0.076 ]
## Significant:  NO
##
##  o3-GPT-Image vs Gemini-2.5
## ---------------------------------------
## Proportions:  0.553  vs  0.458
## Difference:  0.095
## Chi-squared:  11.096
## Degrees of freedom:  1
## P-value:  0.0008651
## 95% CI: [ 0.038 ,  0.151 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Gemini-2.0-Flash
## ---------------------------------------
## Proportions:  0.553  vs  0.395
## Difference:  0.158
## Chi-squared:  31.431
## Degrees of freedom:  1
## P-value:  0.00000002066
## 95% CI: [ 0.103 ,  0.214 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.553  vs  0.324
## Difference:  0.229
## Chi-squared:  39.238
## Degrees of freedom:  1
## P-value:  0.0000000003752
## 95% CI: [ 0.159 ,  0.298 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Sonnet-4
## ---------------------------------------
## Proportions:  0.553  vs  0.421
## Difference:  0.132
## Chi-squared:  21.786
## Degrees of freedom:  1
## P-value:  0.000003049
## 95% CI: [ 0.076 ,  0.188 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Opus-4.1
## ---------------------------------------
```

```
## Proportions:  0.553  vs  0.495
## Difference:  0.058
## Chi-squared:  2.338
## Degrees of freedom:  1
## P-value:  0.1262
## 95% CI: [ -0.016 ,  0.131 ]
## Significant:  NO
##
##  o3-GPT-Image vs GPT-5
## ---------------------------------------
## Proportions:  0.553  vs  0.643
## Difference:  -0.09
## Chi-squared:  10.329
## Degrees of freedom:  1
## P-value:  0.001309
## 95% CI: [ -0.145 ,  -0.035 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-4.1
## ---------------------------------------
## Proportions:  0.635  vs  0.419
## Difference:  0.216
## Chi-squared:  53.314
## Degrees of freedom:  1
## P-value:  0.0000000000002842
## 95% CI: [ 0.158 ,  0.274 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-4.1-GPT-Image
## ---------------------------------------
## Proportions:  0.635  vs  0.398
## Difference:  0.237
## Chi-squared:  64.425
## Degrees of freedom:  1
## P-value:  0.000000000000001003
## 95% CI: [ 0.18 ,  0.295 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs ChatGPT-4o
## ---------------------------------------
## Proportions:  0.635  vs  0.431
## Difference:  0.204
## Chi-squared:  47.79
## Degrees of freedom:  1
## P-value:  0.000000000004744
## 95% CI: [ 0.146 ,  0.263 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o4-mini
## ---------------------------------------
## Proportions:  0.635  vs  0.533
## Difference:  0.102
## Chi-squared:  11.996
## Degrees of freedom:  1
```

```
## P-value:  0.0005333
## 95% CI: [ 0.043 ,  0.161 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.635  vs  0.458
## Difference:  0.177
## Chi-squared:  35.896
## Degrees of freedom:  1
## P-value:  0.000000002081
## 95% CI: [ 0.118 ,  0.235 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.635  vs  0.395
## Difference:  0.24
## Chi-squared:  66.056
## Degrees of freedom:  1
## P-value:  0.0000000000000004382
## 95% CI: [ 0.183 ,  0.298 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.635  vs  0.324
## Difference:  0.311
## Chi-squared:  69.231
## Degrees of freedom:  1
## P-value:  0.00000000000000008757
## 95% CI: [ 0.239 ,  0.382 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Sonnet-4
## ----------------------------------------
## Proportions:  0.635  vs  0.421
## Difference:  0.214
## Chi-squared:  52.518
## Degrees of freedom:  1
## P-value:  0.0000000000004262
## 95% CI: [ 0.156 ,  0.272 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Opus-4.1
## ----------------------------------------
## Proportions:  0.635  vs  0.495
## Difference:  0.14
## Chi-squared:  14.082
## Degrees of freedom:  1
## P-value:  0.000175
## 95% CI: [ 0.065 ,  0.215 ]
## Significant:  YES (p < 0.05)
##
```

```
##  o3-Pro vs GPT-5
## ----------------------------------------
## Proportions:  0.635  vs  0.643
## Difference:  -0.008
## Chi-squared:  0.048
## Degrees of freedom:  1
## P-value:  0.8264
## 95% CI: [ -0.065 ,  0.049 ]
## Significant:  NO
##
##  GPT-4.1 vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.419  vs  0.398
## Difference:  0.022
## Chi-squared:  0.379
## Degrees of freedom:  1
## P-value:  0.5381
## 95% CI: [ -0.043 ,  0.086 ]
## Significant:  NO
##
##  GPT-4.1 vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.419  vs  0.431
## Difference:  -0.012
## Chi-squared:  0.089
## Degrees of freedom:  1
## P-value:  0.7658
## 95% CI: [ -0.076 ,  0.053 ]
## Significant:  NO
##
##  GPT-4.1 vs o4-mini
## ----------------------------------------
## Proportions:  0.419  vs  0.533
## Difference:  -0.114
## Chi-squared:  12.028
## Degrees of freedom:  1
## P-value:  0.000524
## 95% CI: [ -0.179 ,  -0.049 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.419  vs  0.458
## Difference:  -0.039
## Chi-squared:  1.332
## Degrees of freedom:  1
## P-value:  0.2485
## 95% CI: [ -0.104 ,  0.026 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.419  vs  0.395
## Difference:  0.025
```

```
## Chi-squared:  0.506
## Degrees of freedom:  1
## P-value:  0.4771
## 95% CI: [ -0.04 ,  0.089 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.419  vs  0.324
## Difference:  0.095
## Chi-squared:  5.688
## Degrees of freedom:  1
## P-value:  0.01708
## 95% CI: [ 0.018 ,  0.172 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs Sonnet-4
## ---------------------------------------
## Proportions:  0.419  vs  0.421
## Difference:  -0.002
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.066 ,  0.062 ]
## Significant:  NO
##
##  GPT-4.1 vs Opus-4.1
## ---------------------------------------
## Proportions:  0.419  vs  0.495
## Difference:  -0.076
## Chi-squared:  3.445
## Degrees of freedom:  1
## P-value:  0.06344
## 95% CI: [ -0.156 ,  0.004 ]
## Significant:  NO
##
##  GPT-4.1 vs GPT-5
## ---------------------------------------
## Proportions:  0.419  vs  0.643
## Difference:  -0.224
## Chi-squared:  47.369
## Degrees of freedom:  1
## P-value:  0.000000000005879
## 95% CI: [ -0.287 ,  -0.16 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs ChatGPT-4o
## ---------------------------------------
## Proportions:  0.398  vs  0.431
## Difference:  -0.033
## Chi-squared:  0.958
## Degrees of freedom:  1
## P-value:  0.3276
## 95% CI: [ -0.098 ,  0.031 ]
```

```
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs o4-mini
## ---------------------------------------
## Proportions:  0.398  vs  0.533
## Difference:  -0.136
## Chi-squared:  17.173
## Degrees of freedom:  1
## P-value:  0.00003412
## 95% CI: [ -0.2 ,  -0.071 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs Gemini-2.5
## ---------------------------------------
## Proportions:  0.398  vs  0.458
## Difference:  -0.061
## Chi-squared:  3.365
## Degrees of freedom:  1
## P-value:  0.06659
## 95% CI: [ -0.125 ,  0.004 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash
## ---------------------------------------
## Proportions:  0.398  vs  0.395
## Difference:  0.003
## Chi-squared:  0.001
## Degrees of freedom:  1
## P-value:  0.9766
## 95% CI: [ -0.061 ,  0.067 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.398  vs  0.324
## Difference:  0.073
## Chi-squared:  3.375
## Degrees of freedom:  1
## P-value:  0.06621
## 95% CI: [ -0.003 ,  0.15 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Sonnet-4
## ---------------------------------------
## Proportions:  0.398  vs  0.421
## Difference:  -0.023
## Chi-squared:  0.445
## Degrees of freedom:  1
## P-value:  0.5048
## 95% CI: [ -0.088 ,  0.041 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Opus-4.1
## ---------------------------------------
```

```
## Proportions:  0.398  vs  0.495
## Difference:  -0.098
## Chi-squared:  5.834
## Degrees of freedom:  1
## P-value:  0.01572
## 95% CI: [ -0.178 ,  -0.018 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs GPT-5
## ---------------------------------------
## Proportions:  0.398  vs  0.643
## Difference:  -0.245
## Chi-squared:  56.932
## Degrees of freedom:  1
## P-value:  0.00000000000004513
## 95% CI: [ -0.309 ,  -0.182 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs o4-mini
## ---------------------------------------
## Proportions:  0.431  vs  0.533
## Difference:  -0.102
## Chi-squared:  9.655
## Degrees of freedom:  1
## P-value:  0.001889
## 95% CI: [ -0.167 ,  -0.037 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs Gemini-2.5
## ---------------------------------------
## Proportions:  0.431  vs  0.458
## Difference:  -0.027
## Chi-squared:  0.626
## Degrees of freedom:  1
## P-value:  0.4288
## 95% CI: [ -0.092 ,  0.037 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash
## ---------------------------------------
## Proportions:  0.431  vs  0.395
## Difference:  0.036
## Chi-squared:  1.154
## Degrees of freedom:  1
## P-value:  0.2827
## 95% CI: [ -0.028 ,  0.101 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.431  vs  0.324
## Difference:  0.107
## Chi-squared:  7.163
## Degrees of freedom:  1
```

```
## P-value:  0.007442
## 95% CI: [ 0.029 ,  0.184 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs Sonnet-4
## ----------------------------------------
## Proportions:  0.431  vs  0.421
## Difference:  0.01
## Chi-squared:  0.061
## Degrees of freedom:  1
## P-value:  0.8051
## 95% CI: [ -0.055 ,  0.075 ]
## Significant:  NO
##
##  ChatGPT-4o vs Opus-4.1
## ----------------------------------------
## Proportions:  0.431  vs  0.495
## Difference:  -0.064
## Chi-squared:  2.429
## Degrees of freedom:  1
## P-value:  0.1191
## 95% CI: [ -0.145 ,  0.016 ]
## Significant:  NO
##
##  ChatGPT-4o vs GPT-5
## ----------------------------------------
## Proportions:  0.431  vs  0.643
## Difference:  -0.212
## Chi-squared:  42.614
## Degrees of freedom:  1
## P-value:  0.00000000006669
## 95% CI: [ -0.276 ,  -0.148 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.533  vs  0.458
## Difference:  0.075
## Chi-squared:  5.082
## Degrees of freedom:  1
## P-value:  0.02417
## 95% CI: [ 0.01 ,  0.14 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.533  vs  0.395
## Difference:  0.139
## Chi-squared:  17.965
## Degrees of freedom:  1
## P-value:  0.0000225
## 95% CI: [ 0.074 ,  0.203 ]
## Significant:  YES (p < 0.05)
##
```

```
##  o4-mini vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.533  vs  0.324
## Difference:  0.209
## Chi-squared:  27.238
## Degrees of freedom:  1
## P-value:  0.0000001799
## 95% CI: [ 0.132 ,  0.286 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Sonnet-4
## ---------------------------------------
## Proportions:  0.533  vs  0.421
## Difference:  0.112
## Chi-squared:  11.678
## Degrees of freedom:  1
## P-value:  0.0006324
## 95% CI: [ 0.047 ,  0.177 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Opus-4.1
## ---------------------------------------
## Proportions:  0.533  vs  0.495
## Difference:  0.038
## Chi-squared:  0.773
## Degrees of freedom:  1
## P-value:  0.3793
## 95% CI: [ -0.043 ,  0.118 ]
## Significant:  NO
##
##  o4-mini vs GPT-5
## ---------------------------------------
## Proportions:  0.533  vs  0.643
## Difference:  -0.11
## Chi-squared:  11.515
## Degrees of freedom:  1
## P-value:  0.0006902
## 95% CI: [ -0.174 ,  -0.046 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5 vs Gemini-2.0-Flash
## ---------------------------------------
## Proportions:  0.458  vs  0.395
## Difference:  0.064
## Chi-squared:  3.723
## Degrees of freedom:  1
## P-value:  0.05366
## 95% CI: [ -0.001 ,  0.128 ]
## Significant:  NO
##
##  Gemini-2.5 vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.458  vs  0.324
## Difference:  0.134
```

```
## Chi-squared:  11.305
## Degrees of freedom:  1
## P-value:  0.0007729
## 95% CI: [ 0.057 ,  0.211 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5 vs Sonnet-4
## ---------------------------------------
## Proportions:  0.458  vs  0.421
## Difference:  0.037
## Chi-squared:  1.216
## Degrees of freedom:  1
## P-value:  0.2701
## 95% CI: [ -0.027 ,  0.102 ]
## Significant:  NO
##
##  Gemini-2.5 vs Opus-4.1
## ---------------------------------------
## Proportions:  0.458  vs  0.495
## Difference:  -0.037
## Chi-squared:  0.737
## Degrees of freedom:  1
## P-value:  0.3908
## 95% CI: [ -0.117 ,  0.044 ]
## Significant:  NO
##
##  Gemini-2.5 vs GPT-5
## ---------------------------------------
## Proportions:  0.458  vs  0.643
## Difference:  -0.185
## Chi-squared:  32.359
## Degrees of freedom:  1
## P-value:  0.00000001281
## 95% CI: [ -0.249 ,  -0.121 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.395  vs  0.324
## Difference:  0.07
## Chi-squared:  3.098
## Degrees of freedom:  1
## P-value:  0.07838
## 95% CI: [ -0.006 ,  0.147 ]
## Significant:  NO
##
##  Gemini-2.0-Flash vs Sonnet-4
## ---------------------------------------
## Proportions:  0.395  vs  0.421
## Difference:  -0.026
## Chi-squared:  0.581
## Degrees of freedom:  1
## P-value:  0.4459
## 95% CI: [ -0.09 ,  0.038 ]
```

```
## Significant:  NO
##
##  Gemini-2.0-Flash vs Opus-4.1
## ----------------------------------------
## Proportions:  0.395  vs  0.495
## Difference:  -0.101
## Chi-squared:  6.218
## Degrees of freedom:  1
## P-value:  0.01264
## 95% CI: [ -0.181 ,  -0.021 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs GPT-5
## ----------------------------------------
## Proportions:  0.395  vs  0.643
## Difference:  -0.248
## Chi-squared:  58.335
## Degrees of freedom:  1
## P-value:  0.0000000000000221
## 95% CI: [ -0.312 ,  -0.185 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-GPT-Image vs Sonnet-4
## ----------------------------------------
## Proportions:  0.324  vs  0.421
## Difference:  -0.097
## Chi-squared:  5.886
## Degrees of freedom:  1
## P-value:  0.01526
## 95% CI: [ -0.174 ,  -0.02 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-GPT-Image vs Opus-4.1
## ----------------------------------------
## Proportions:  0.324  vs  0.495
## Difference:  -0.171
## Chi-squared:  13.814
## Degrees of freedom:  1
## P-value:  0.0002018
## 95% CI: [ -0.262 ,  -0.08 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-GPT-Image vs GPT-5
## ----------------------------------------
## Proportions:  0.324  vs  0.643
## Difference:  -0.319
## Chi-squared:  64.109
## Degrees of freedom:  1
## P-value:  0.000000000000001177
## 95% CI: [ -0.395 ,  -0.243 ]
## Significant:  YES (p < 0.05)
##
##  Sonnet-4 vs Opus-4.1
## ----------------------------------------
```

```
## Proportions:  0.421  vs  0.495
## Difference:  -0.074
## Chi-squared:  3.291
## Degrees of freedom:  1
## P-value:  0.06966
## 95% CI: [ -0.155 ,  0.006 ]
## Significant:  NO
##
##  Sonnet-4 vs GPT-5
## ----------------------------------------
## Proportions:  0.421  vs  0.643
## Difference:  -0.222
## Chi-squared:  46.684
## Degrees of freedom:  1
## P-value:  0.00000000000834
## 95% CI: [ -0.286 ,  -0.159 ]
## Significant:  YES (p < 0.05)
##
##  Opus-4.1 vs GPT-5
## ----------------------------------------
## Proportions:  0.495  vs  0.643
## Difference:  -0.148
## Chi-squared:  13.873
## Degrees of freedom:  1
## P-value:  0.0001956
## 95% CI: [ -0.227 ,  -0.068 ]
## 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  sig
## ------------  ---------------------------------------------  -------  ------------  --------  ----
## X-squared      Humans vs o3                                    -0.137    47.9063736    0.0000  TRU
## X-squared1     Humans vs o3-GPT-Image                          -0.033     3.5265237    0.0604  FALS
## X-squared2     Humans vs o3-Pro                                -0.115    33.9167412    0.0000  TRU
## X-squared3     Humans vs GPT-4.1                                0.100    17.5291122    0.0000  TRU
## X-squared4     Humans vs GPT-4.1-GPT-Image                      0.122    25.9899741    0.0000  TRU
## X-squared5     Humans vs ChatGPT-4o                             0.089    13.6769284    0.0002  TRU
## X-squared6     Humans vs o4-mini                               -0.013     0.2714854    0.6023  FALS
## X-squared7     Humans vs Gemini-2.5                             0.061     6.4544305    0.0111  TRU
## X-squared8     Humans vs Gemini-2.0-Flash                       0.125    27.2999900    0.0000  TRU
```

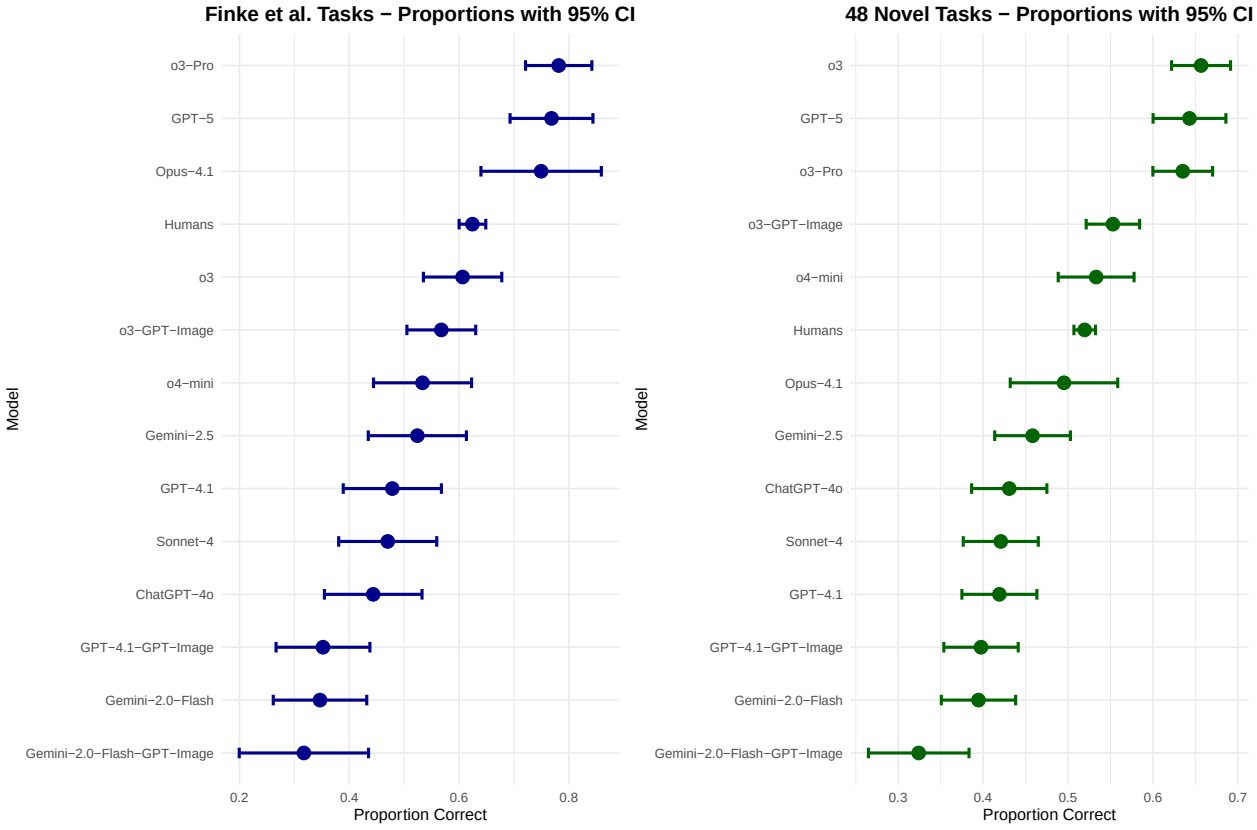
```
## X-squared9    Humans vs Gemini-2.0-Flash-GPT-Image              0.195    34.4775625    0.0000  TRU
## X-squared10   Humans vs Sonnet-4                               0.099    16.9578299    0.0000  TRU
## X-squared11   Humans vs Opus-4.1                               0.024     0.4557580    0.4996  FALS
## X-squared12   Humans vs GPT-5                                 -0.123    26.6438423    0.0000  TRU
## X-squared13   o3 vs o3-GPT-Image                              0.104    17.9907647    0.0000  TRU
## X-squared14   o3 vs o3-Pro                                    0.022     0.6407169    0.4235  FALS
## X-squared15   o3 vs GPT-4.1                                   0.237    64.9699894    0.0000  TRU
## X-squared16   o3 vs GPT-4.1-GPT-Image                         0.259    77.1264773    0.0000  TRU
## X-squared17   o3 vs ChatGPT-4o                               0.226    58.8787355    0.0000  TRU
## X-squared18   o3 vs o4-mini                                   0.124    17.9073744    0.0000  TRU
## X-squared19   o3 vs Gemini-2.5                               0.198    45.6250283    0.0000  TRU
## X-squared20   o3 vs Gemini-2.0-Flash                         0.262    78.9018995    0.0000  TRU
## X-squared21   o3 vs Gemini-2.0-Flash-GPT-Image               0.332    79.9497081    0.0000  TRU
## X-squared22   o3 vs Sonnet-4                                 0.236    64.0943686    0.0000  TRU
## X-squared23   o3 vs Opus-4.1                                 0.161    19.1427553    0.0000  TRU
## X-squared24   o3 vs GPT-5                                    0.014     0.1790592    0.6722  FALS
## X-squared25   o3-GPT-Image vs o3-Pro                        -0.082    11.1407366    0.0008  TRU
## X-squared26   o3-GPT-Image vs GPT-4.1                        0.134    22.3332196    0.0000  TRU
## X-squared27   o3-GPT-Image vs GPT-4.1-GPT-Image              0.155    30.2351997    0.0000  TRU
## X-squared28   o3-GPT-Image vs ChatGPT-4o                     0.122    18.5914296    0.0000  TRU
## X-squared29   o3-GPT-Image vs o4-mini                        0.020     0.4267414    0.5136  FALS
## X-squared30   o3-GPT-Image vs Gemini-2.5                     0.095    11.0960806    0.0009  TRU
## X-squared31   o3-GPT-Image vs Gemini-2.0-Flash               0.158    31.4313142    0.0000  TRU
## X-squared32   o3-GPT-Image vs Gemini-2.0-Flash-GPT-Image     0.229    39.2379561    0.0000  TRU
## X-squared33   o3-GPT-Image vs Sonnet-4                       0.132    21.7855202    0.0000  TRU
## X-squared34   o3-GPT-Image vs Opus-4.1                       0.058     2.3382046    0.1262  FALS
## X-squared35   o3-GPT-Image vs GPT-5                         -0.090    10.3293789    0.0013  TRU
## X-squared36   o3-Pro vs GPT-4.1                             0.216    53.3144149    0.0000  TRU
## X-squared37   o3-Pro vs GPT-4.1-GPT-Image                   0.237    64.4253856    0.0000  TRU
## X-squared38   o3-Pro vs ChatGPT-4o                          0.204    47.7899487    0.0000  TRU
## X-squared39   o3-Pro vs o4-mini                             0.102    11.9955058    0.0005  TRU
## X-squared40   o3-Pro vs Gemini-2.5                          0.177    35.8960130    0.0000  TRU
## X-squared41   o3-Pro vs Gemini-2.0-Flash                    0.240    66.0564477    0.0000  TRU
## X-squared42   o3-Pro vs Gemini-2.0-Flash-GPT-Image          0.311    69.2312173    0.0000  TRU
## X-squared43   o3-Pro vs Sonnet-4                            0.214    52.5183409    0.0000  TRU
## X-squared44   o3-Pro vs Opus-4.1                            0.140    14.0821041    0.0002  TRU
## X-squared45   o3-Pro vs GPT-5                              -0.008     0.0480925    0.8264  FALS
## X-squared46   GPT-4.1 vs GPT-4.1-GPT-Image                  0.022     0.3791307    0.5381  FALS
## X-squared47   GPT-4.1 vs ChatGPT-4o                        -0.012     0.0887562    0.7658  FALS
## X-squared48   GPT-4.1 vs o4-mini                          -0.114    12.0284272    0.0005  TRU
## X-squared49   GPT-4.1 vs Gemini-2.5                       -0.039     1.3318985    0.2485  FALS
## X-squared50   GPT-4.1 vs Gemini-2.0-Flash                  0.025     0.5055811    0.4771  FALS
## X-squared51   GPT-4.1 vs Gemini-2.0-Flash-GPT-Image        0.095     5.6877759    0.0171  TRU
## X-squared52   GPT-4.1 vs Sonnet-4                         -0.002     0.0000000    1.0000  FALS
## X-squared53   GPT-4.1 vs Opus-4.1                         -0.076     3.4451895    0.0634  FALS
## X-squared54   GPT-4.1 vs GPT-5                           -0.224    47.3694148    0.0000  TRU
## X-squared55   GPT-4.1-GPT-Image vs ChatGPT-4o             -0.033     0.9583187    0.3276  FALS
## X-squared56   GPT-4.1-GPT-Image vs o4-mini               -0.136    17.1734561    0.0000  TRU
## X-squared57   GPT-4.1-GPT-Image vs Gemini-2.5            -0.061     3.3650488    0.0666  FALS
## X-squared58   GPT-4.1-GPT-Image vs Gemini-2.0-Flash       0.003     0.0008633    0.9766  FALS
## X-squared59   GPT-4.1-GPT-Image vs Gemini-2.0-Flash-GPT-Image  0.073  3.3745012    0.0662  FALS
## X-squared60   GPT-4.1-GPT-Image vs Sonnet-4               -0.023     0.4448430    0.5048  FALS
## X-squared61   GPT-4.1-GPT-Image vs Opus-4.1               -0.098     5.8337201    0.0157  TRU
## X-squared62   GPT-4.1-GPT-Image vs GPT-5                  -0.245    56.9315098    0.0000  TRU
```

```
## X-squared63   ChatGPT-4o vs o4-mini                                -0.102     9.6545783    0.0019  TRU
## X-squared64   ChatGPT-4o vs Gemini-2.5                            -0.027     0.6259449    0.4288  FALS
## X-squared65   ChatGPT-4o vs Gemini-2.0-Flash                       0.036     1.1539281    0.2827  FALS
## X-squared66   ChatGPT-4o vs Gemini-2.0-Flash-GPT-Image            0.107     7.1629741    0.0074  TRU
## X-squared67   ChatGPT-4o vs Sonnet-4                              0.010     0.0608684    0.8051  FALS
## X-squared68   ChatGPT-4o vs Opus-4.1                             -0.064     2.4291860    0.1191  FALS
## X-squared69   ChatGPT-4o vs GPT-5                                -0.212    42.6136662    0.0000  TRU
## X-squared70   o4-mini vs Gemini-2.5                              0.075     5.0823538    0.0242  TRU
## X-squared71   o4-mini vs Gemini-2.0-Flash                         0.139    17.9649751    0.0000  TRU
## X-squared72   o4-mini vs Gemini-2.0-Flash-GPT-Image              0.209    27.2378408    0.0000  TRU
## X-squared73   o4-mini vs Sonnet-4                                0.112    11.6780639    0.0006  TRU
## X-squared74   o4-mini vs Opus-4.1                               0.038     0.7730084    0.3793  FALS
## X-squared75   o4-mini vs GPT-5                                  -0.110    11.5154242    0.0007  TRU
## X-squared76   Gemini-2.5 vs Gemini-2.0-Flash                     0.064     3.7231515    0.0537  FALS
## X-squared77   Gemini-2.5 vs Gemini-2.0-Flash-GPT-Image          0.134    11.3051198    0.0008  TRU
## X-squared78   Gemini-2.5 vs Sonnet-4                            0.037     1.2164067    0.2701  FALS
## X-squared79   Gemini-2.5 vs Opus-4.1                           -0.037     0.7365986    0.3908  FALS
## X-squared80   Gemini-2.5 vs GPT-5                              -0.185    32.3592466    0.0000  TRU
## X-squared81   Gemini-2.0-Flash vs Gemini-2.0-Flash-GPT-Image    0.070     3.0981827    0.0784  FALS
## X-squared82   Gemini-2.0-Flash vs Sonnet-4                     -0.026     0.5810526    0.4459  FALS
## X-squared83   Gemini-2.0-Flash vs Opus-4.1                     -0.101     6.2184578    0.0126  TRU
## X-squared84   Gemini-2.0-Flash vs GPT-5                        -0.248    58.3354535    0.0000  TRU
## X-squared85   Gemini-2.0-Flash-GPT-Image vs Sonnet-4           -0.097     5.8856350    0.0153  TRU
## X-squared86   Gemini-2.0-Flash-GPT-Image vs Opus-4.1           -0.171    13.8139982    0.0002  TRU
## X-squared87   Gemini-2.0-Flash-GPT-Image vs GPT-5              -0.319    64.1088048    0.0000  TRU
## X-squared88   Sonnet-4 vs Opus-4.1                            -0.074     3.2911089    0.0697  FALS
## X-squared89   Sonnet-4 vs GPT-5                               -0.222    46.6842465    0.0000  TRU
## X-squared90   Opus-4.1 vs GPT-5                               -0.148    13.8730259    0.0002  TRU
```

Visualization of All Comparisons

```
# Plot 1: Proportions with confidence intervals for Finke tasks
finke_plot <- ggplot(finke_data, aes(x = reorder(model, proportion), y = proportion)) +
  geom_point(size = 4, color = "darkblue") +
  geom_errorbar(aes(ymin = proportion - 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    ymax = proportion + 1.96 * sqrt(proportion * (1 - proportion) / max_score)),
                width = 0.2, size = 1, color = "darkblue") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Finke et al. Tasks - Proportions with 95% CI",
       x = "Model",
       y = "Proportion Correct") +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))

# Plot 2: Proportions with confidence intervals for 48 Novel tasks
novel_48_plot <- ggplot(novel_data, aes(x = reorder(model, proportion), y = proportion)) +
  geom_point(size = 4, color = "darkgreen") +
  geom_errorbar(aes(ymin = proportion - 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    ymax = proportion + 1.96 * sqrt(proportion * (1 - proportion) / max_score)),
                width = 0.2, size = 1, color = "darkgreen") +
  coord_flip() +
  theme_minimal() +
  labs(title = "48 Novel Tasks - Proportions with 95% CI",
       x = "Model",
```

```
       y = "Proportion Correct") +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))

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

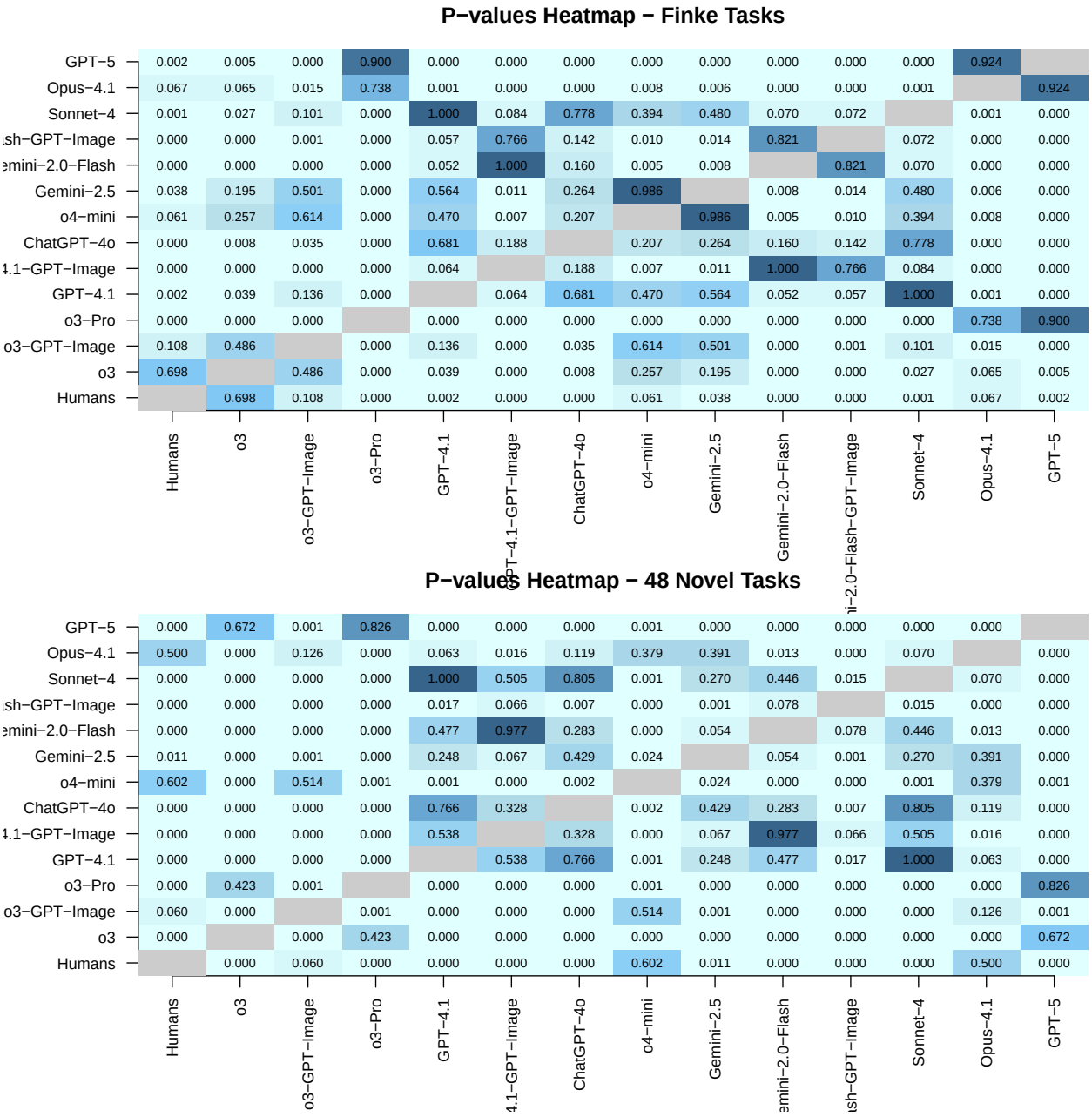


Heatmap of P-values

```
# 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
```

```r
novel_models <- novel_data$model
novel_pval_matrix <- matrix(NA, nrow = length(novel_models), ncol = length(novel_models))
rownames(novel_pval_matrix) <- novel_models
colnames(novel_pval_matrix) <- novel_models

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

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

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

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

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

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

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

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

```
     las = 2, cex.axis = 0.8)

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

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



Summary of Significant Differences

```
# 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:
```

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

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

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

```
## Finke Tasks:
```

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

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

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

```
##   Significant differences: 54
```

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

```
##   Percentage significant: 59.3 %
```

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

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

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

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

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

```
##   Significant differences: 61
```

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

```
##   Percentage significant: 67 %
```

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

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

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

```
##
##
##                comparison                                  diff   p_value
## -------------  ------------------------------------------  --------  --------
## X-squared2     Humans vs o3-Pro                            -0.1571    0.0000
## X-squared3     Humans vs GPT-4.1                            0.1459    0.0022
## X-squared4     Humans vs GPT-4.1-GPT-Image                  0.2722    0.0000
## X-squared5     Humans vs ChatGPT-4o                         0.1807    0.0001
## X-squared7     Humans vs Gemini-2.5                         0.1004    0.0375
## X-squared8     Humans vs Gemini-2.0-Flash                   0.2777    0.0000
## X-squared9     Humans vs Gemini-2.0-Flash-GPT-Image         0.3070    0.0000
## X-squared10    Humans vs Sonnet-4                           0.1543    0.0012
## X-squared12    Humans vs GPT-5                             -0.1440    0.0022
## X-squared14    o3 vs o3-Pro                                -0.1750    0.0005
## X-squared15    o3 vs GPT-4.1                                0.1280    0.0387
## X-squared16    o3 vs GPT-4.1-GPT-Image                      0.2543    0.0000
## X-squared17    o3 vs ChatGPT-4o                             0.1628    0.0080
```

```
## X-squared20   o3 vs Gemini-2.0-Flash                          0.2597    0.0000
## X-squared21   o3 vs Gemini-2.0-Flash-GPT-Image               0.2891    0.0002
## X-squared22   o3 vs Sonnet-4                                  0.1364    0.0272
## X-squared24   o3 vs GPT-5                                    -0.1619    0.0051
## X-squared25   o3-GPT-Image vs o3-Pro                         -0.2138    0.0000
## X-squared27   o3-GPT-Image vs GPT-4.1-GPT-Image               0.2156    0.0002
## X-squared28   o3-GPT-Image vs ChatGPT-4o                      0.1240    0.0348
## X-squared31   o3-GPT-Image vs Gemini-2.0-Flash                0.2210    0.0001
## X-squared32   o3-GPT-Image vs Gemini-2.0-Flash-GPT-Image      0.2503    0.0009
## X-squared34   o3-GPT-Image vs Opus-4.1                       -0.1818    0.0152
## X-squared35   o3-GPT-Image vs GPT-5                          -0.2007    0.0003
## X-squared36   o3-Pro vs GPT-4.1                               0.3030    0.0000
## X-squared37   o3-Pro vs GPT-4.1-GPT-Image                     0.4294    0.0000
## X-squared38   o3-Pro vs ChatGPT-4o                            0.3378    0.0000
## X-squared39   o3-Pro vs o4-mini                               0.2480    0.0000
## X-squared40   o3-Pro vs Gemini-2.5                            0.2575    0.0000
## X-squared41   o3-Pro vs Gemini-2.0-Flash                      0.4348    0.0000
## X-squared42   o3-Pro vs Gemini-2.0-Flash-GPT-Image           0.4641    0.0000
## X-squared43   o3-Pro vs Sonnet-4                              0.3114    0.0000
## X-squared53   GPT-4.1 vs Opus-4.1                            -0.2711    0.0010
## X-squared54   GPT-4.1 vs GPT-5                               -0.2899    0.0000
## X-squared56   GPT-4.1-GPT-Image vs o4-mini                   -0.1813    0.0070
## X-squared57   GPT-4.1-GPT-Image vs Gemini-2.5                -0.1719    0.0107
## X-squared61   GPT-4.1-GPT-Image vs Opus-4.1                  -0.3974    0.0000
## X-squared62   GPT-4.1-GPT-Image vs GPT-5                     -0.4162    0.0000
## X-squared68   ChatGPT-4o vs Opus-4.1                         -0.3059    0.0002
## X-squared69   ChatGPT-4o vs GPT-5                            -0.3247    0.0000
## X-squared71   o4-mini vs Gemini-2.0-Flash                     0.1867    0.0054
## X-squared72   o4-mini vs Gemini-2.0-Flash-GPT-Image          0.2161    0.0098
## X-squared74   o4-mini vs Opus-4.1                            -0.2160    0.0084
## X-squared75   o4-mini vs GPT-5                               -0.2349    0.0002
## X-squared76   Gemini-2.5 vs Gemini-2.0-Flash                  0.1773    0.0083
## X-squared77   Gemini-2.5 vs Gemini-2.0-Flash-GPT-Image       0.2067    0.0137
## X-squared79   Gemini-2.5 vs Opus-4.1                         -0.2255    0.0060
## X-squared80   Gemini-2.5 vs GPT-5                            -0.2444    0.0001
## X-squared83   Gemini-2.0-Flash vs Opus-4.1                   -0.4028    0.0000
## X-squared84   Gemini-2.0-Flash vs GPT-5                      -0.4217    0.0000
## X-squared86   Gemini-2.0-Flash-GPT-Image vs Opus-4.1         -0.4321    0.0000
## X-squared87   Gemini-2.0-Flash-GPT-Image vs GPT-5            -0.4510    0.0000
## X-squared88   Sonnet-4 vs Opus-4.1                          -0.2794    0.0007
## X-squared89   Sonnet-4 vs GPT-5                             -0.2983    0.0000
```

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

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

```
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                              -0.1370    0.0000
## X-squared2    Humans vs o3-Pro                          -0.1154    0.0000
## X-squared3    Humans vs GPT-4.1                          0.1004    0.0000
## X-squared4    Humans vs GPT-4.1-GPT-Image                0.1220    0.0000
## X-squared5    Humans vs ChatGPT-4o                       0.0888    0.0002
## X-squared7    Humans vs Gemini-2.5                       0.0614    0.0111
## X-squared8    Humans vs Gemini-2.0-Flash                 0.1250    0.0000
## X-squared9    Humans vs Gemini-2.0-Flash-GPT-Image       0.1954    0.0000
## X-squared10   Humans vs Sonnet-4                         0.0988    0.0000
## X-squared12   Humans vs GPT-5                           -0.1234    0.0000
## X-squared13   o3 vs o3-GPT-Image                         0.1038    0.0000
## X-squared15   o3 vs GPT-4.1                             0.2374    0.0000
## X-squared16   o3 vs GPT-4.1-GPT-Image                   0.2590    0.0000
## X-squared17   o3 vs ChatGPT-4o                          0.2258    0.0000
## X-squared18   o3 vs o4-mini                            0.1235    0.0000
## X-squared19   o3 vs Gemini-2.5                         0.1984    0.0000
## X-squared20   o3 vs Gemini-2.0-Flash                   0.2620    0.0000
## X-squared21   o3 vs Gemini-2.0-Flash-GPT-Image         0.3324    0.0000
## X-squared22   o3 vs Sonnet-4                           0.2358    0.0000
## X-squared23   o3 vs Opus-4.1                           0.1614    0.0000
## X-squared25   o3-GPT-Image vs o3-Pro                  -0.0822    0.0008
## X-squared26   o3-GPT-Image vs GPT-4.1                   0.1336    0.0000
## X-squared27   o3-GPT-Image vs GPT-4.1-GPT-Image         0.1553    0.0000
## X-squared28   o3-GPT-Image vs ChatGPT-4o                0.1220    0.0000
## X-squared30   o3-GPT-Image vs Gemini-2.5                0.0946    0.0009
## X-squared31   o3-GPT-Image vs Gemini-2.0-Flash          0.1583    0.0000
## X-squared32   o3-GPT-Image vs Gemini-2.0-Flash-GPT-Image  0.2286  0.0000
## X-squared33   o3-GPT-Image vs Sonnet-4                  0.1320    0.0000
## X-squared35   o3-GPT-Image vs GPT-5                    -0.0902    0.0013
## X-squared36   o3-Pro vs GPT-4.1                         0.2158    0.0000
## X-squared37   o3-Pro vs GPT-4.1-GPT-Image               0.2375    0.0000
## X-squared38   o3-Pro vs ChatGPT-4o                      0.2043    0.0000
## X-squared39   o3-Pro vs o4-mini                        0.1020    0.0005
## X-squared40   o3-Pro vs Gemini-2.5                     0.1768    0.0000
## X-squared41   o3-Pro vs Gemini-2.0-Flash               0.2405    0.0000
## X-squared42   o3-Pro vs Gemini-2.0-Flash-GPT-Image     0.3108    0.0000
## X-squared43   o3-Pro vs Sonnet-4                       0.2142    0.0000
## X-squared44   o3-Pro vs Opus-4.1                       0.1398    0.0002
## X-squared48   GPT-4.1 vs o4-mini                      -0.1139    0.0005
## X-squared51   GPT-4.1 vs Gemini-2.0-Flash-GPT-Image    0.0950    0.0171
## X-squared54   GPT-4.1 vs GPT-5                        -0.2238    0.0000
## X-squared56   GPT-4.1-GPT-Image vs o4-mini            -0.1355    0.0000
## X-squared61   GPT-4.1-GPT-Image vs Opus-4.1           -0.0977    0.0157
## X-squared62   GPT-4.1-GPT-Image vs GPT-5             -0.2454    0.0000
## X-squared63   ChatGPT-4o vs o4-mini                   -0.1023    0.0019
## X-squared66   ChatGPT-4o vs Gemini-2.0-Flash-GPT-Image  0.1066   0.0074
## X-squared69   ChatGPT-4o vs GPT-5                     -0.2122    0.0000
## X-squared70   o4-mini vs Gemini-2.5                    0.0748    0.0242
## X-squared71   o4-mini vs Gemini-2.0-Flash              0.1385    0.0000
## X-squared72   o4-mini vs Gemini-2.0-Flash-GPT-Image    0.2089    0.0000
## X-squared73   o4-mini vs Sonnet-4                      0.1123    0.0006
```

```
## X-squared75   o4-mini vs GPT-5                                -0.1099    0.0007
## X-squared77   Gemini-2.5 vs Gemini-2.0-Flash-GPT-Image       0.1340    0.0008
## X-squared80   Gemini-2.5 vs GPT-5                           -0.1847    0.0000
## X-squared83   Gemini-2.0-Flash vs Opus-4.1                  -0.1007    0.0126
## X-squared84   Gemini-2.0-Flash vs GPT-5                     -0.2484    0.0000
## X-squared85   Gemini-2.0-Flash-GPT-Image vs Sonnet-4        -0.0966    0.0153
## X-squared86   Gemini-2.0-Flash-GPT-Image vs Opus-4.1        -0.1710    0.0002
## X-squared87   Gemini-2.0-Flash-GPT-Image vs GPT-5           -0.3188    0.0000
## X-squared89   Sonnet-4 vs GPT-5                            -0.2222    0.0000
## X-squared90   Opus-4.1 vs GPT-5                            -0.1477    0.0002
```

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:  0.541  vs  0.647
## Difference:  -0.106
## Chi-squared:  35.828
## Degrees of freedom:  1
## P-value:  0.000000002156
## 95% CI: [ -0.139 ,  -0.072 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o3-GPT-Image
## ----------------------------------------
## Proportions:  0.541  vs  0.556
## Difference:  -0.015
## Chi-squared:  0.861
```

```
## Degrees of freedom:  1
## P-value:  0.3533
## 95% CI: [ -0.046 ,  0.016 ]
## Significant:  NO
##
##  Humans vs o3-Pro
## ----------------------------------------
## Proportions:  0.541  vs  0.664
## Difference:  -0.123
## Chi-squared:  49.009
## Degrees of freedom:  1
## P-value:  0.000000000002548
## 95% CI: [ -0.157 ,  -0.09 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-4.1
## ----------------------------------------
## Proportions:  0.541  vs  0.431
## Difference:  0.11
## Chi-squared:  26.512
## Degrees of freedom:  1
## P-value:  0.0000002619
## 95% CI: [ 0.068 ,  0.152 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.541  vs  0.388
## Difference:  0.152
## Chi-squared:  51.216
## Degrees of freedom:  1
## P-value:  0.0000000000008273
## 95% CI: [ 0.111 ,  0.194 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.541  vs  0.433
## Difference:  0.108
## Chi-squared:  25.401
## Degrees of freedom:  1
## P-value:  0.0000004658
## 95% CI: [ 0.065 ,  0.15 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o4-mini
## ----------------------------------------
## Proportions:  0.541  vs  0.533
## Difference:  0.008
## Chi-squared:  0.106
## Degrees of freedom:  1
## P-value:  0.7448
## 95% CI: [ -0.035 ,  0.05 ]
## Significant:  NO
```

```
##
##  Humans vs Gemini-2.5
## ---------------------------------------
## Proportions:  0.541  vs  0.471
## Difference:  0.07
## Chi-squared:  10.524
## Degrees of freedom:  1
## P-value:  0.001178
## 95% CI: [ 0.027 ,  0.112 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash
## ---------------------------------------
## Proportions:  0.541  vs  0.385
## Difference:  0.156
## Chi-squared:  53.6
## Degrees of freedom:  1
## P-value:  0.0000000000002457
## 95% CI: [ 0.115 ,  0.197 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.541  vs  0.323
## Difference:  0.218
## Chi-squared:  54.243
## Degrees of freedom:  1
## P-value:  0.0000000000001772
## 95% CI: [ 0.162 ,  0.274 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Sonnet-4
## ---------------------------------------
## Proportions:  0.541  vs  0.431
## Difference:  0.11
## Chi-squared:  26.692
## Degrees of freedom:  1
## P-value:  0.0000002387
## 95% CI: [ 0.068 ,  0.152 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs Opus-4.1
## ---------------------------------------
## Proportions:  0.541  vs  0.546
## Difference:  -0.005
## Chi-squared:  0.014
## Degrees of freedom:  1
## P-value:  0.9074
## 95% CI: [ -0.064 ,  0.054 ]
## Significant:  NO
##
##  Humans vs GPT-5
## ---------------------------------------
## Proportions:  0.541  vs  0.668
```

```
## Difference:  -0.127
## Chi-squared:  35.763
## Degrees of freedom:  1
## P-value:  0.000000002229
## 95% CI: [ -0.167 ,  -0.087 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o3-GPT-Image
## ----------------------------------------
## Proportions:  0.647  vs  0.556
## Difference:  0.091
## Chi-squared:  17.204
## Degrees of freedom:  1
## P-value:  0.00003357
## 95% CI: [ 0.048 ,  0.134 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o3-Pro
## ----------------------------------------
## Proportions:  0.647  vs  0.664
## Difference:  -0.018
## Chi-squared:  0.552
## Degrees of freedom:  1
## P-value:  0.4575
## 95% CI: [ -0.063 ,  0.027 ]
## Significant:  NO
##
##  o3 vs GPT-4.1
## ----------------------------------------
## Proportions:  0.647  vs  0.431
## Difference:  0.216
## Chi-squared:  67.006
## Degrees of freedom:  1
## P-value:  0.0000000000000002707
## 95% CI: [ 0.164 ,  0.267 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.647  vs  0.388
## Difference:  0.258
## Chi-squared:  95.601
## Degrees of freedom:  1
## P-value:  0.00000000000000000000001405
## 95% CI: [ 0.207 ,  0.309 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.647  vs  0.433
## Difference:  0.213
## Chi-squared:  65.597
## Degrees of freedom:  1
## P-value:  0.0000000000000005532
```

```
## 95% CI: [ 0.161 ,  0.265 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs o4-mini
## ---------------------------------------
## Proportions:  0.647  vs  0.533
## Difference:  0.113
## Chi-squared:  18.842
## Degrees of freedom:  1
## P-value:  0.0000142
## 95% CI: [ 0.061 ,  0.165 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.5
## ---------------------------------------
## Proportions:  0.647  vs  0.471
## Difference:  0.175
## Chi-squared:  44.531
## Degrees of freedom:  1
## P-value:  0.00000000002504
## 95% CI: [ 0.123 ,  0.227 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.0-Flash
## ---------------------------------------
## Proportions:  0.647  vs  0.385
## Difference:  0.262
## Chi-squared:  98.173
## Degrees of freedom:  1
## P-value:  0.00000000000000000000003834
## 95% CI: [ 0.21 ,  0.313 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.647  vs  0.323
## Difference:  0.324
## Chi-squared:  94.659
## Degrees of freedom:  1
## P-value:  0.0000000000000000000002262
## 95% CI: [ 0.26 ,  0.387 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Sonnet-4
## ---------------------------------------
## Proportions:  0.647  vs  0.431
## Difference:  0.216
## Chi-squared:  67.232
## Degrees of freedom:  1
## P-value:  0.0000000000000002413
## 95% CI: [ 0.164 ,  0.268 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs Opus-4.1
```

```
## ----------------------------------------
## Proportions:  0.647  vs  0.546
## Difference:  0.1
## Chi-squared:  9.234
## Degrees of freedom:  1
## P-value:  0.002376
## 95% CI: [ 0.034 ,  0.167 ]
## Significant:  YES (p < 0.05)
##
##  o3 vs GPT-5
## ----------------------------------------
## Proportions:  0.647  vs  0.668
## Difference:  -0.021
## Chi-squared:  0.644
## Degrees of freedom:  1
## P-value:  0.4223
## 95% CI: [ -0.072 ,  0.029 ]
## Significant:  NO
##
##  o3-GPT-Image vs o3-Pro
## ----------------------------------------
## Proportions:  0.556  vs  0.664
## Difference:  -0.109
## Chi-squared:  24.836
## Degrees of freedom:  1
## P-value:  0.0000006244
## 95% CI: [ -0.151 ,  -0.066 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs GPT-4.1
## ----------------------------------------
## Proportions:  0.556  vs  0.431
## Difference:  0.125
## Chi-squared:  24.426
## Degrees of freedom:  1
## P-value:  0.0000007722
## 95% CI: [ 0.075 ,  0.175 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.556  vs  0.388
## Difference:  0.167
## Chi-squared:  44.124
## Degrees of freedom:  1
## P-value:  0.00000000003082
## 95% CI: [ 0.118 ,  0.217 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.556  vs  0.433
## Difference:  0.122
## Chi-squared:  23.522
```

```
## Degrees of freedom:  1
## P-value:  0.000001235
## 95% CI: [ 0.073 ,  0.172 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs o4-mini
## ----------------------------------------
## Proportions:  0.556  vs  0.533
## Difference:  0.023
## Chi-squared:  0.739
## Degrees of freedom:  1
## P-value:  0.39
## 95% CI: [ -0.027 ,  0.073 ]
## Significant:  NO
##
##  o3-GPT-Image vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.556  vs  0.471
## Difference:  0.084
## Chi-squared:  11.097
## Degrees of freedom:  1
## P-value:  0.0008646
## 95% CI: [ 0.034 ,  0.134 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.556  vs  0.385
## Difference:  0.171
## Chi-squared:  45.999
## Degrees of freedom:  1
## P-value:  0.00000000001183
## 95% CI: [ 0.122 ,  0.22 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.556  vs  0.323
## Difference:  0.233
## Chi-squared:  51.189
## Degrees of freedom:  1
## P-value:  0.0000000000008389
## 95% CI: [ 0.171 ,  0.295 ]
## Significant:  YES (p < 0.05)
##
##  o3-GPT-Image vs Sonnet-4
## ----------------------------------------
## Proportions:  0.556  vs  0.431
## Difference:  0.125
## Chi-squared:  24.572
## Degrees of freedom:  1
## P-value:  0.0000007159
## 95% CI: [ 0.075 ,  0.175 ]
## Significant:  YES (p < 0.05)
```

```
##
##  o3-GPT-Image vs Opus-4.1
## ----------------------------------------
## Proportions:  0.556  vs  0.546
## Difference:  0.01
## Chi-squared:  0.057
## Degrees of freedom:  1
## P-value:  0.812
## 95% CI: [ -0.055 ,  0.075 ]
## Significant:  NO
##
##  o3-GPT-Image vs GPT-5
## ----------------------------------------
## Proportions:  0.556  vs  0.668
## Difference:  -0.112
## Chi-squared:  20.427
## Degrees of freedom:  1
## P-value:  0.000006194
## 95% CI: [ -0.161 ,  -0.064 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-4.1
## ----------------------------------------
## Proportions:  0.664  vs  0.431
## Difference:  0.233
## Chi-squared:  79.028
## Degrees of freedom:  1
## P-value:  0.00000000000000000006124
## 95% CI: [ 0.182 ,  0.285 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-4.1-GPT-Image
## ----------------------------------------
## Proportions:  0.664  vs  0.388
## Difference:  0.276
## Chi-squared:  109.744
## Degrees of freedom:  1
## P-value:  0.00000000000000000000000001115
## 95% CI: [ 0.225 ,  0.327 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.664  vs  0.433
## Difference:  0.231
## Chi-squared:  77.502
## Degrees of freedom:  1
## P-value:  0.00000000000000001326
## 95% CI: [ 0.179 ,  0.283 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs o4-mini
## ----------------------------------------
## Proportions:  0.664  vs  0.533
```

```
## Difference:  0.131
## Chi-squared:  25.53
## Degrees of freedom:  1
## P-value:  0.0000004355
## 95% CI: [ 0.079 ,  0.183 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.664  vs  0.471
## Difference:  0.193
## Chi-squared:  54.484
## Degrees of freedom:  1
## P-value:  0.0000000000001567
## 95% CI: [ 0.141 ,  0.245 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.664  vs  0.385
## Difference:  0.279
## Chi-squared:  112.486
## Degrees of freedom:  1
## P-value:  0.00000000000000000000000002796
## 95% CI: [ 0.228 ,  0.33 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.664  vs  0.323
## Difference:  0.341
## Chi-squared:  106.236
## Degrees of freedom:  1
## P-value:  0.00000000000000000000000006545
## 95% CI: [ 0.278 ,  0.405 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Sonnet-4
## ----------------------------------------
## Proportions:  0.664  vs  0.431
## Difference:  0.234
## Chi-squared:  79.273
## Degrees of freedom:  1
## P-value:  0.000000000000000005409
## 95% CI: [ 0.182 ,  0.285 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs Opus-4.1
## ----------------------------------------
## Proportions:  0.664  vs  0.546
## Difference:  0.118
## Chi-squared:  13.062
## Degrees of freedom:  1
## P-value:  0.0003013
```

```
## 95% CI: [ 0.052 ,  0.185 ]
## Significant:  YES (p < 0.05)
##
##  o3-Pro vs GPT-5
## ---------------------------------------
## Proportions:  0.664  vs  0.668
## Difference:  -0.004
## Chi-squared:  0.009
## Degrees of freedom:  1
## P-value:  0.9248
## 95% CI: [ -0.054 ,  0.046 ]
## Significant:  NO
##
##  GPT-4.1 vs GPT-4.1-GPT-Image
## ---------------------------------------
## Proportions:  0.431  vs  0.388
## Difference:  0.043
## Chi-squared:  2.074
## Degrees of freedom:  1
## P-value:  0.1498
## 95% CI: [ -0.015 ,  0.1 ]
## Significant:  NO
##
##  GPT-4.1 vs ChatGPT-4o
## ---------------------------------------
## Proportions:  0.431  vs  0.433
## Difference:  -0.002
## Chi-squared:  0.001
## Degrees of freedom:  1
## P-value:  0.982
## 95% CI: [ -0.06 ,  0.055 ]
## Significant:  NO
##
##  GPT-4.1 vs o4-mini
## ---------------------------------------
## Proportions:  0.431  vs  0.533
## Difference:  -0.102
## Chi-squared:  12.123
## Degrees of freedom:  1
## P-value:  0.0004979
## 95% CI: [ -0.16 ,  -0.044 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs Gemini-2.5
## ---------------------------------------
## Proportions:  0.431  vs  0.471
## Difference:  -0.04
## Chi-squared:  1.813
## Degrees of freedom:  1
## P-value:  0.1781
## 95% CI: [ -0.098 ,  0.018 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.0-Flash
```

```
## ----------------------------------------
## Proportions:  0.431  vs  0.385
## Difference:  0.046
## Chi-squared:  2.447
## Degrees of freedom:  1
## P-value:  0.1178
## 95% CI: [ -0.011 ,  0.103 ]
## Significant:  NO
##
##  GPT-4.1 vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.431  vs  0.323
## Difference:  0.108
## Chi-squared:  9.351
## Degrees of freedom:  1
## P-value:  0.002228
## 95% CI: [ 0.04 ,  0.177 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs Sonnet-4
## ----------------------------------------
## Proportions:  0.431  vs  0.431
## Difference:  0
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.056 ,  0.057 ]
## Significant:  NO
##
##  GPT-4.1 vs Opus-4.1
## ----------------------------------------
## Proportions:  0.431  vs  0.546
## Difference:  -0.115
## Chi-squared:  10.171
## Degrees of freedom:  1
## P-value:  0.001427
## 95% CI: [ -0.186 ,  -0.044 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1 vs GPT-5
## ----------------------------------------
## Proportions:  0.431  vs  0.668
## Difference:  -0.237
## Chi-squared:  67.127
## Degrees of freedom:  1
## P-value:  0.0000000000000002546
## 95% CI: [ -0.293 ,  -0.181 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs ChatGPT-4o
## ----------------------------------------
## Proportions:  0.388  vs  0.433
## Difference:  -0.045
## Chi-squared:  2.313
```

```
## Degrees of freedom:  1
## P-value:  0.1283
## 95% CI: [ -0.102 ,  0.012 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs o4-mini
## ----------------------------------------
## Proportions:  0.388  vs  0.533
## Difference:  -0.145
## Chi-squared:  24.693
## Degrees of freedom:  1
## P-value:  0.0000006724
## 95% CI: [ -0.202 ,  -0.087 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.388  vs  0.471
## Difference:  -0.083
## Chi-squared:  8.079
## Degrees of freedom:  1
## P-value:  0.004477
## 95% CI: [ -0.14 ,  -0.025 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.388  vs  0.385
## Difference:  0.003
## Chi-squared:  0.004
## Degrees of freedom:  1
## P-value:  0.9482
## 95% CI: [ -0.053 ,  0.06 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.388  vs  0.323
## Difference:  0.066
## Chi-squared:  3.434
## Degrees of freedom:  1
## P-value:  0.06385
## 95% CI: [ -0.003 ,  0.134 ]
## Significant:  NO
##
##  GPT-4.1-GPT-Image vs Sonnet-4
## ----------------------------------------
## Proportions:  0.388  vs  0.431
## Difference:  -0.042
## Chi-squared:  2.037
## Degrees of freedom:  1
## P-value:  0.1535
## 95% CI: [ -0.099 ,  0.015 ]
## Significant:  NO
```

```
##
##  GPT-4.1-GPT-Image vs Opus-4.1
## ---------------------------------------
## Proportions:  0.388  vs  0.546
## Difference:  -0.158
## Chi-squared:  19.517
## Degrees of freedom:  1
## P-value:  0.000009971
## 95% CI: [ -0.229 ,  -0.087 ]
## Significant:  YES (p < 0.05)
##
##  GPT-4.1-GPT-Image vs GPT-5
## ---------------------------------------
## Proportions:  0.388  vs  0.668
## Difference:  -0.28
## Chi-squared:  92.977
## Degrees of freedom:  1
## P-value:  0.000000000000000000005291
## 95% CI: [ -0.335 ,  -0.224 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs o4-mini
## ---------------------------------------
## Proportions:  0.433  vs  0.533
## Difference:  -0.1
## Chi-squared:  11.57
## Degrees of freedom:  1
## P-value:  0.0006702
## 95% CI: [ -0.158 ,  -0.042 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs Gemini-2.5
## ---------------------------------------
## Proportions:  0.433  vs  0.471
## Difference:  -0.038
## Chi-squared:  1.602
## Degrees of freedom:  1
## P-value:  0.2056
## 95% CI: [ -0.096 ,  0.02 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash
## ---------------------------------------
## Proportions:  0.433  vs  0.385
## Difference:  0.048
## Chi-squared:  2.706
## Degrees of freedom:  1
## P-value:  0.09999
## 95% CI: [ -0.009 ,  0.106 ]
## Significant:  NO
##
##  ChatGPT-4o vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.433  vs  0.323
```

```
## Difference:  0.111
## Chi-squared:  9.751
## Degrees of freedom:  1
## P-value:  0.001792
## 95% CI: [ 0.042 ,  0.179 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs Sonnet-4
## ----------------------------------------
## Proportions:  0.433  vs  0.431
## Difference:  0.003
## Chi-squared:  0.001
## Degrees of freedom:  1
## P-value:  0.9717
## 95% CI: [ -0.055 ,  0.06 ]
## Significant:  NO
##
##  ChatGPT-4o vs Opus-4.1
## ----------------------------------------
## Proportions:  0.433  vs  0.546
## Difference:  -0.113
## Chi-squared:  9.754
## Degrees of freedom:  1
## P-value:  0.001789
## 95% CI: [ -0.184 ,  -0.041 ]
## Significant:  YES (p < 0.05)
##
##  ChatGPT-4o vs GPT-5
## ----------------------------------------
## Proportions:  0.433  vs  0.668
## Difference:  -0.235
## Chi-squared:  65.846
## Degrees of freedom:  1
## P-value:  0.0000000000000004877
## 95% CI: [ -0.291 ,  -0.178 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Gemini-2.5
## ----------------------------------------
## Proportions:  0.533  vs  0.471
## Difference:  0.062
## Chi-squared:  4.334
## Degrees of freedom:  1
## P-value:  0.03736
## 95% CI: [ 0.004 ,  0.12 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Gemini-2.0-Flash
## ----------------------------------------
## Proportions:  0.533  vs  0.385
## Difference:  0.148
## Chi-squared:  25.928
## Degrees of freedom:  1
## P-value:  0.0000003544
```

```
## 95% CI: [ 0.091 ,  0.206 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Gemini-2.0-Flash-GPT-Image
## ---------------------------------------
## Proportions:  0.533  vs  0.323
## Difference:  0.21
## Chi-squared:  34.74
## Degrees of freedom:  1
## P-value:  0.000000003767
## 95% CI: [ 0.142 ,  0.279 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Sonnet-4
## ---------------------------------------
## Proportions:  0.533  vs  0.431
## Difference:  0.102
## Chi-squared:  12.213
## Degrees of freedom:  1
## P-value:  0.0004746
## 95% CI: [ 0.045 ,  0.16 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini vs Opus-4.1
## ---------------------------------------
## Proportions:  0.533  vs  0.546
## Difference:  -0.013
## Chi-squared:  0.087
## Degrees of freedom:  1
## P-value:  0.7674
## 95% CI: [ -0.084 ,  0.059 ]
## Significant:  NO
##
##  o4-mini vs GPT-5
## ---------------------------------------
## Proportions:  0.533  vs  0.668
## Difference:  -0.135
## Chi-squared:  22.201
## Degrees of freedom:  1
## P-value:  0.000002456
## 95% CI: [ -0.191 ,  -0.078 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5 vs Gemini-2.0-Flash
## ---------------------------------------
## Proportions:  0.471  vs  0.385
## Difference:  0.086
## Chi-squared:  8.798
## Degrees of freedom:  1
## P-value:  0.003016
## 95% CI: [ 0.029 ,  0.144 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5 vs Gemini-2.0-Flash-GPT-Image
```

```
## ----------------------------------------
## Proportions:  0.471  vs  0.323
## Difference:  0.149
## Chi-squared:  17.493
## Degrees of freedom:  1
## P-value:  0.00002884
## 95% CI: [ 0.08 ,  0.217 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5 vs Sonnet-4
## ----------------------------------------
## Proportions:  0.471  vs  0.431
## Difference:  0.041
## Chi-squared:  1.848
## Degrees of freedom:  1
## P-value:  0.174
## 95% CI: [ -0.017 ,  0.099 ]
## Significant:  NO
##
##  Gemini-2.5 vs Opus-4.1
## ----------------------------------------
## Proportions:  0.471  vs  0.546
## Difference:  -0.075
## Chi-squared:  4.169
## Degrees of freedom:  1
## P-value:  0.04116
## 95% CI: [ -0.146 ,  -0.003 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.5 vs GPT-5
## ----------------------------------------
## Proportions:  0.471  vs  0.668
## Difference:  -0.197
## Chi-squared:  46.533
## Degrees of freedom:  1
## P-value:  0.000000000009007
## 95% CI: [ -0.253 ,  -0.14 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs Gemini-2.0-Flash-GPT-Image
## ----------------------------------------
## Proportions:  0.385  vs  0.323
## Difference:  0.062
## Chi-squared:  3.073
## Degrees of freedom:  1
## P-value:  0.07959
## 95% CI: [ -0.006 ,  0.13 ]
## Significant:  NO
##
##  Gemini-2.0-Flash vs Sonnet-4
## ----------------------------------------
## Proportions:  0.385  vs  0.431
## Difference:  -0.046
## Chi-squared:  2.406
```

```
## Degrees of freedom:  1
## P-value:  0.1208
## 95% CI: [ -0.103 ,  0.012 ]
## Significant:  NO
##
##  Gemini-2.0-Flash vs Opus-4.1
## ----------------------------------------
## Proportions:  0.385  vs  0.546
## Difference:  -0.161
## Chi-squared:  20.429
## Degrees of freedom:  1
## P-value:  0.000006189
## 95% CI: [ -0.232 ,  -0.09 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash vs GPT-5
## ----------------------------------------
## Proportions:  0.385  vs  0.668
## Difference:  -0.283
## Chi-squared:  95.291
## Degrees of freedom:  1
## P-value:  0.000000000000000000001643
## 95% CI: [ -0.339 ,  -0.227 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-GPT-Image vs Sonnet-4
## ----------------------------------------
## Proportions:  0.323  vs  0.431
## Difference:  -0.108
## Chi-squared:  9.288
## Degrees of freedom:  1
## P-value:  0.002307
## 95% CI: [ -0.176 ,  -0.039 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-GPT-Image vs Opus-4.1
## ----------------------------------------
## Proportions:  0.323  vs  0.546
## Difference:  -0.223
## Chi-squared:  29.525
## Degrees of freedom:  1
## P-value:  0.00000005521
## 95% CI: [ -0.304 ,  -0.143 ]
## Significant:  YES (p < 0.05)
##
##  Gemini-2.0-Flash-GPT-Image vs GPT-5
## ----------------------------------------
## Proportions:  0.323  vs  0.668
## Difference:  -0.345
## Chi-squared:  95.032
## Degrees of freedom:  1
## P-value:  0.000000000000000000001873
## 95% CI: [ -0.413 ,  -0.278 ]
## Significant:  YES (p < 0.05)
```

```
##
##  Sonnet-4 vs Opus-4.1
## ----------------------------------------
## Proportions:  0.431  vs  0.546
## Difference:  -0.115
## Chi-squared:  10.238
## Degrees of freedom:  1
## P-value:  0.001375
## 95% CI: [ -0.187 ,  -0.044 ]
## Significant:  YES (p < 0.05)
##
##  Sonnet-4 vs GPT-5
## ----------------------------------------
## Proportions:  0.431  vs  0.668
## Difference:  -0.237
## Chi-squared:  67.333
## Degrees of freedom:  1
## P-value:  0.0000000000000002294
## 95% CI: [ -0.294 ,  -0.181 ]
## Significant:  YES (p < 0.05)
##
##  Opus-4.1 vs GPT-5
## ----------------------------------------
## Proportions:  0.546  vs  0.668
## Difference:  -0.122
## Chi-squared:  12.212
## Degrees of freedom:  1
## P-value:  0.0004749
## 95% CI: [ -0.192 ,  -0.052 ]
## Significant:  YES (p < 0.05)
```

```r
# 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:
```

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

```
##
##
##                comparison                                          diff    chi_squared   p_value  sig
## -------------  --------------------------------------------  -------  ------------  --------  ----
## X-squared      Humans vs o3                                     -0.106    35.8276131    0.0000  TRU
## X-squared1     Humans vs o3-GPT-Image                           -0.015     0.8614579    0.3533  FALS
## X-squared2     Humans vs o3-Pro                                 -0.123    49.0092591    0.0000  TRU
## X-squared3     Humans vs GPT-4.1                                 0.110    26.5119398    0.0000  TRU
## X-squared4     Humans vs GPT-4.1-GPT-Image                       0.152    51.2161943    0.0000  TRU
## X-squared5     Humans vs ChatGPT-4o                              0.108    25.4005296    0.0000  TRU
```

```
## X-squared6    Humans vs o4-mini                                0.008     0.1059207    0.7448  FALS
## X-squared7    Humans vs Gemini-2.5                             0.070    10.5238006    0.0012  TRUE
## X-squared8    Humans vs Gemini-2.0-Flash                       0.156    53.6003119    0.0000  TRUE
## X-squared9    Humans vs Gemini-2.0-Flash-GPT-Image             0.218    54.2430365    0.0000  TRUE
## X-squared10   Humans vs Sonnet-4                               0.110    26.6915827    0.0000  TRUE
## X-squared11   Humans vs Opus-4.1                              -0.005     0.0135338    0.9074  FALS
## X-squared12   Humans vs GPT-5                                 -0.127    35.7628507    0.0000  TRUE
## X-squared13   o3 vs o3-GPT-Image                              0.091    17.2042549    0.0000  TRUE
## X-squared14   o3 vs o3-Pro                                   -0.018     0.5520756    0.4575  FALS
## X-squared15   o3 vs GPT-4.1                                   0.216    67.0059236    0.0000  TRUE
## X-squared16   o3 vs GPT-4.1-GPT-Image                        0.258    95.6014083    0.0000  TRUE
## X-squared17   o3 vs ChatGPT-4o                               0.213    65.5970668    0.0000  TRUE
## X-squared18   o3 vs o4-mini                                  0.113    18.8417640    0.0000  TRUE
## X-squared19   o3 vs Gemini-2.5                               0.175    44.5305665    0.0000  TRUE
## X-squared20   o3 vs Gemini-2.0-Flash                         0.262    98.1731228    0.0000  TRUE
## X-squared21   o3 vs Gemini-2.0-Flash-GPT-Image               0.324    94.6591471    0.0000  TRUE
## X-squared22   o3 vs Sonnet-4                                 0.216    67.2322266    0.0000  TRUE
## X-squared23   o3 vs Opus-4.1                                 0.100     9.2339743    0.0024  TRUE
## X-squared24   o3 vs GPT-5                                   -0.021     0.6439982    0.4223  FALS
## X-squared25   o3-GPT-Image vs o3-Pro                        -0.109    24.8355052    0.0000  TRUE
## X-squared26   o3-GPT-Image vs GPT-4.1                        0.125    24.4260210    0.0000  TRUE
## X-squared27   o3-GPT-Image vs GPT-4.1-GPT-Image              0.167    44.1238606    0.0000  TRUE
## X-squared28   o3-GPT-Image vs ChatGPT-4o                     0.122    23.5223961    0.0000  TRUE
## X-squared29   o3-GPT-Image vs o4-mini                        0.023     0.7390554    0.3900  FALS
## X-squared30   o3-GPT-Image vs Gemini-2.5                     0.084    11.0971962    0.0009  TRUE
## X-squared31   o3-GPT-Image vs Gemini-2.0-Flash               0.171    45.9993236    0.0000  TRUE
## X-squared32   o3-GPT-Image vs Gemini-2.0-Flash-GPT-Image     0.233    51.1888485    0.0000  TRUE
## X-squared33   o3-GPT-Image vs Sonnet-4                       0.125    24.5718705    0.0000  TRUE
## X-squared34   o3-GPT-Image vs Opus-4.1                       0.010     0.0565431    0.8120  FALS
## X-squared35   o3-GPT-Image vs GPT-5                         -0.112    20.4271734    0.0000  TRUE
## X-squared36   o3-Pro vs GPT-4.1                             0.233    79.0280100    0.0000  TRUE
## X-squared37   o3-Pro vs GPT-4.1-GPT-Image                   0.276   109.7435810    0.0000  TRUE
## X-squared38   o3-Pro vs ChatGPT-4o                          0.231    77.5021635    0.0000  TRUE
## X-squared39   o3-Pro vs o4-mini                             0.131    25.5300958    0.0000  TRUE
## X-squared40   o3-Pro vs Gemini-2.5                          0.193    54.4844383    0.0000  TRUE
## X-squared41   o3-Pro vs Gemini-2.0-Flash                    0.279   112.4863221    0.0000  TRUE
## X-squared42   o3-Pro vs Gemini-2.0-Flash-GPT-Image          0.341   106.2361959    0.0000  TRUE
## X-squared43   o3-Pro vs Sonnet-4                            0.234    79.2729742    0.0000  TRUE
## X-squared44   o3-Pro vs Opus-4.1                            0.118    13.0623225    0.0003  TRUE
## X-squared45   o3-Pro vs GPT-5                              -0.004     0.0089137    0.9248  FALS
## X-squared46   GPT-4.1 vs GPT-4.1-GPT-Image                  0.043     2.0740631    0.1498  FALS
## X-squared47   GPT-4.1 vs ChatGPT-4o                        -0.002     0.0005083    0.9820  FALS
## X-squared48   GPT-4.1 vs o4-mini                           -0.102    12.1234375    0.0005  TRUE
## X-squared49   GPT-4.1 vs Gemini-2.5                        -0.040     1.8132502    0.1781  FALS
## X-squared50   GPT-4.1 vs Gemini-2.0-Flash                  0.046     2.4466570    0.1178  FALS
## X-squared51   GPT-4.1 vs Gemini-2.0-Flash-GPT-Image        0.108     9.3511547    0.0022  TRUE
## X-squared52   GPT-4.1 vs Sonnet-4                          0.000     0.0000000    1.0000  FALS
## X-squared53   GPT-4.1 vs Opus-4.1                         -0.115    10.1710341    0.0014  TRUE
## X-squared54   GPT-4.1 vs GPT-5                           -0.237    67.1268951    0.0000  TRUE
## X-squared55   GPT-4.1-GPT-Image vs ChatGPT-4o             -0.045     2.3131962    0.1283  FALS
## X-squared56   GPT-4.1-GPT-Image vs o4-mini                -0.145    24.6926284    0.0000  TRUE
## X-squared57   GPT-4.1-GPT-Image vs Gemini-2.5             -0.083     8.0794442    0.0045  TRUE
## X-squared58   GPT-4.1-GPT-Image vs Gemini-2.0-Flash        0.003     0.0042148    0.9482  FALS
## X-squared59   GPT-4.1-GPT-Image vs Gemini-2.0-Flash-GPT-Image  0.066     3.4343525    0.0639  FALS
```

```
## X-squared60   GPT-4.1-GPT-Image vs Sonnet-4                    -0.042     2.0370752    0.1535  FALS
## X-squared61   GPT-4.1-GPT-Image vs Opus-4.1                    -0.158    19.5168932    0.0000  TRUE
## X-squared62   GPT-4.1-GPT-Image vs GPT-5                       -0.280    92.9767197    0.0000  TRUE
## X-squared63   ChatGPT-4o vs o4-mini                           -0.100    11.5700331    0.0007  TRUE
## X-squared64   ChatGPT-4o vs Gemini-2.5                        -0.038     1.6023198    0.2056  FALS
## X-squared65   ChatGPT-4o vs Gemini-2.0-Flash                   0.048     2.7057747    0.1000  FALS
## X-squared66   ChatGPT-4o vs Gemini-2.0-Flash-GPT-Image        0.111     9.7514954    0.0018  TRUE
## X-squared67   ChatGPT-4o vs Sonnet-4                          0.003     0.0012568    0.9717  FALS
## X-squared68   ChatGPT-4o vs Opus-4.1                         -0.113     9.7539193    0.0018  TRUE
## X-squared69   ChatGPT-4o vs GPT-5                            -0.235    65.8456009    0.0000  TRUE
## X-squared70   o4-mini vs Gemini-2.5                          0.062     4.3338132    0.0374  TRUE
## X-squared71   o4-mini vs Gemini-2.0-Flash                     0.148    25.9278947    0.0000  TRUE
## X-squared72   o4-mini vs Gemini-2.0-Flash-GPT-Image          0.210    34.7404168    0.0000  TRUE
## X-squared73   o4-mini vs Sonnet-4                            0.102    12.2130094    0.0005  TRUE
## X-squared74   o4-mini vs Opus-4.1                           -0.013     0.0874948    0.7674  FALS
## X-squared75   o4-mini vs GPT-5                              -0.135    22.2007055    0.0000  TRUE
## X-squared76   Gemini-2.5 vs Gemini-2.0-Flash                 0.086     8.7978130    0.0030  TRUE
## X-squared77   Gemini-2.5 vs Gemini-2.0-Flash-GPT-Image      0.149    17.4930816    0.0000  TRUE
## X-squared78   Gemini-2.5 vs Sonnet-4                        0.041     1.8481447    0.1740  FALS
## X-squared79   Gemini-2.5 vs Opus-4.1                       -0.075     4.1694039    0.0412  TRUE
## X-squared80   Gemini-2.5 vs GPT-5                          -0.197    46.5334655    0.0000  TRUE
## X-squared81   Gemini-2.0-Flash vs Gemini-2.0-Flash-GPT-Image  0.062     3.0732937    0.0796  FALS
## X-squared82   Gemini-2.0-Flash vs Sonnet-4                  -0.046     2.4064765    0.1208  FALS
## X-squared83   Gemini-2.0-Flash vs Opus-4.1                 -0.161    20.4288379    0.0000  TRUE
## X-squared84   Gemini-2.0-Flash vs GPT-5                    -0.283    95.2914418    0.0000  TRUE
## X-squared85   Gemini-2.0-Flash-GPT-Image vs Sonnet-4       -0.108     9.2879832    0.0023  TRUE
## X-squared86   Gemini-2.0-Flash-GPT-Image vs Opus-4.1       -0.223    29.5245490    0.0000  TRUE
## X-squared87   Gemini-2.0-Flash-GPT-Image vs GPT-5          -0.345    95.0320711    0.0000  TRUE
## X-squared88   Sonnet-4 vs Opus-4.1                         -0.115    10.2384703    0.0014  TRUE
## X-squared89   Sonnet-4 vs GPT-5                            -0.237    67.3326173    0.0000  TRUE
## X-squared90   Opus-4.1 vs GPT-5                           -0.122    12.2118454    0.0005  TRUE
```

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

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

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

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

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

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

```
##   Significant differences: 67
```

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

```
##   Percentage significant: 73.6 %
```

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

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

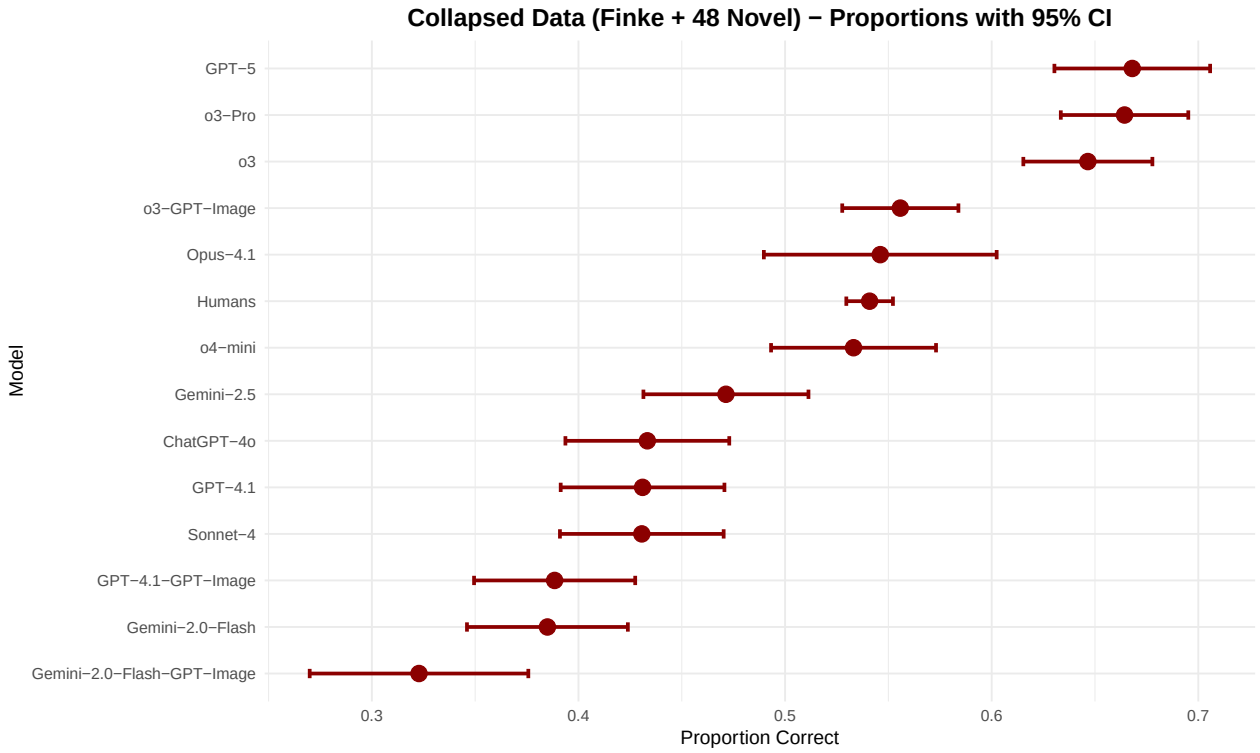
```r
collapsed_sig <- collapsed_results[collapsed_results$significant, c("comparison", "diff", "p_value")]
if (nrow(collapsed_sig) > 0) {
  print(kable(collapsed_sig, format = "simple", digits = 4))
} else {
  cat("  None\n")
}
```

```
##
##
##                comparison                                  diff   p_value
## ------------  ----------------------------------------  --------  --------
## X-squared     Humans vs o3                              -0.1056    0.0000
## X-squared2    Humans vs o3-Pro                          -0.1234    0.0000
## X-squared3    Humans vs GPT-4.1                          0.1099    0.0000
## X-squared4    Humans vs GPT-4.1-GPT-Image                0.1525    0.0000
## X-squared5    Humans vs ChatGPT-4o                       0.1076    0.0000
## X-squared7    Humans vs Gemini-2.5                       0.0695    0.0012
## X-squared8    Humans vs Gemini-2.0-Flash                 0.1559    0.0000
## X-squared9    Humans vs Gemini-2.0-Flash-GPT-Image       0.2181    0.0000
## X-squared10   Humans vs Sonnet-4                         0.1103    0.0000
## X-squared12   Humans vs GPT-5                           -0.1271    0.0000
## X-squared13   o3 vs o3-GPT-Image                         0.0908    0.0000
## X-squared15   o3 vs GPT-4.1                             0.2155    0.0000
## X-squared16   o3 vs GPT-4.1-GPT-Image                   0.2581    0.0000
## X-squared17   o3 vs ChatGPT-4o                          0.2132    0.0000
## X-squared18   o3 vs o4-mini                            0.1134    0.0000
## X-squared19   o3 vs Gemini-2.5                         0.1752    0.0000
## X-squared20   o3 vs Gemini-2.0-Flash                   0.2616    0.0000
## X-squared21   o3 vs Gemini-2.0-Flash-GPT-Image         0.3237    0.0000
## X-squared22   o3 vs Sonnet-4                           0.2159    0.0000
## X-squared23   o3 vs Opus-4.1                           0.1005    0.0024
## X-squared25   o3-GPT-Image vs o3-Pro                   -0.1085    0.0000
## X-squared26   o3-GPT-Image vs GPT-4.1                   0.1248    0.0000
## X-squared27   o3-GPT-Image vs GPT-4.1-GPT-Image        0.1673    0.0000
## X-squared28   o3-GPT-Image vs ChatGPT-4o                0.1224    0.0000
## X-squared30   o3-GPT-Image vs Gemini-2.5                0.0844    0.0009
## X-squared31   o3-GPT-Image vs Gemini-2.0-Flash          0.1708    0.0000
## X-squared32   o3-GPT-Image vs Gemini-2.0-Flash-GPT-Image  0.2330    0.0000
## X-squared33   o3-GPT-Image vs Sonnet-4                  0.1251    0.0000
## X-squared35   o3-GPT-Image vs GPT-5                    -0.1123    0.0000
## X-squared36   o3-Pro vs GPT-4.1                        0.2333    0.0000
## X-squared37   o3-Pro vs GPT-4.1-GPT-Image              0.2758    0.0000
## X-squared38   o3-Pro vs ChatGPT-4o                     0.2310    0.0000
## X-squared39   o3-Pro vs o4-mini                        0.1312    0.0000
## X-squared40   o3-Pro vs Gemini-2.5                     0.1929    0.0000
## X-squared41   o3-Pro vs Gemini-2.0-Flash               0.2793    0.0000
## X-squared42   o3-Pro vs Gemini-2.0-Flash-GPT-Image     0.3415    0.0000
## X-squared43   o3-Pro vs Sonnet-4                       0.2337    0.0000
## X-squared44   o3-Pro vs Opus-4.1                       0.1182    0.0003
## X-squared48   GPT-4.1 vs o4-mini                      -0.1021    0.0005
## X-squared51   GPT-4.1 vs Gemini-2.0-Flash-GPT-Image    0.1082    0.0022
## X-squared53   GPT-4.1 vs Opus-4.1                     -0.1150    0.0014
## X-squared54   GPT-4.1 vs GPT-5                       -0.2370    0.0000
## X-squared56   GPT-4.1-GPT-Image vs o4-mini            -0.1447    0.0000
```

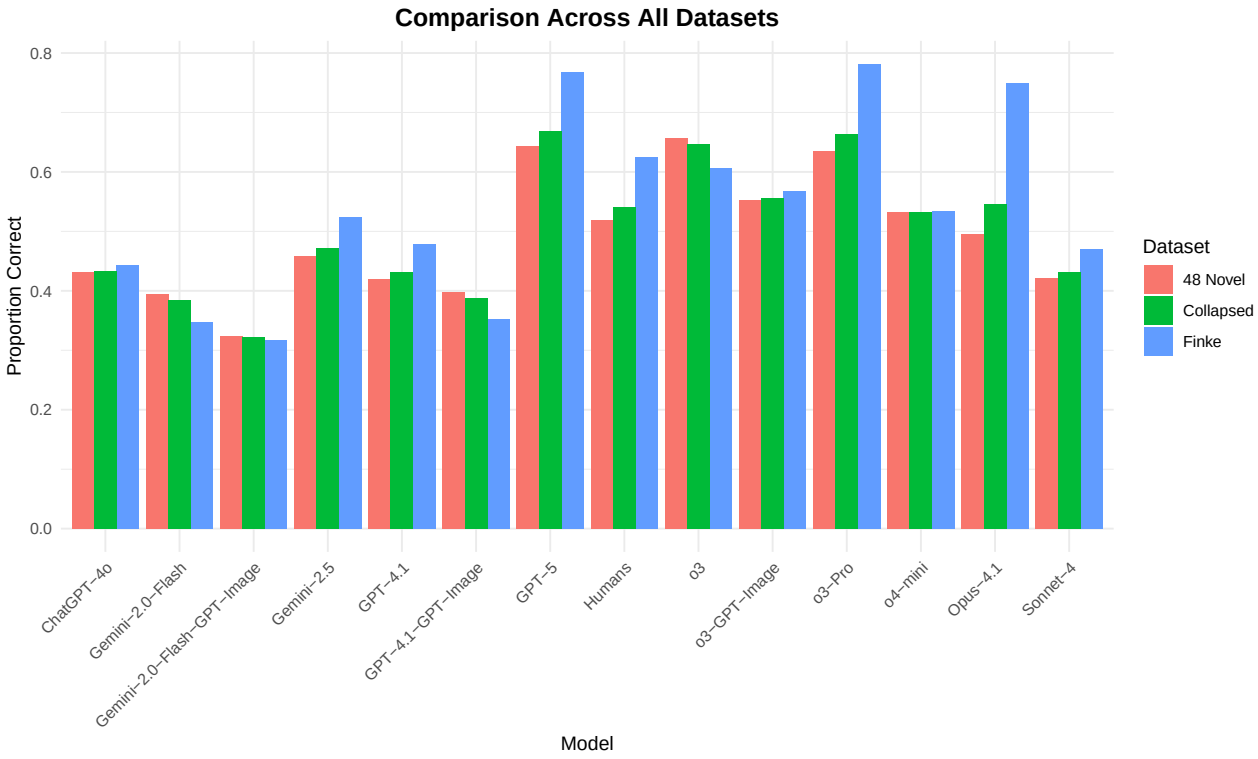
```
## X-squared57   GPT-4.1-GPT-Image vs Gemini-2.5              -0.0829    0.0045
## X-squared61   GPT-4.1-GPT-Image vs Opus-4.1               -0.1576    0.0000
## X-squared62   GPT-4.1-GPT-Image vs GPT-5                  -0.2796    0.0000
## X-squared63   ChatGPT-4o vs o4-mini                       -0.0998    0.0007
## X-squared66   ChatGPT-4o vs Gemini-2.0-Flash-GPT-Image     0.1105    0.0018
## X-squared68   ChatGPT-4o vs Opus-4.1                      -0.1127    0.0018
## X-squared69   ChatGPT-4o vs GPT-5                         -0.2347    0.0000
## X-squared70   o4-mini vs Gemini-2.5                       0.0618    0.0374
## X-squared71   o4-mini vs Gemini-2.0-Flash                 0.1482    0.0000
## X-squared72   o4-mini vs Gemini-2.0-Flash-GPT-Image       0.2103    0.0000
## X-squared73   o4-mini vs Sonnet-4                         0.1025    0.0005
## X-squared75   o4-mini vs GPT-5                           -0.1349    0.0000
## X-squared76   Gemini-2.5 vs Gemini-2.0-Flash              0.0864    0.0030
## X-squared77   Gemini-2.5 vs Gemini-2.0-Flash-GPT-Image    0.1486    0.0000
## X-squared79   Gemini-2.5 vs Opus-4.1                      -0.0747    0.0412
## X-squared80   Gemini-2.5 vs GPT-5                         -0.1967    0.0000
## X-squared83   Gemini-2.0-Flash vs Opus-4.1                -0.1611    0.0000
## X-squared84   Gemini-2.0-Flash vs GPT-5                   -0.2831    0.0000
## X-squared85   Gemini-2.0-Flash-GPT-Image vs Sonnet-4      -0.1078    0.0023
## X-squared86   Gemini-2.0-Flash-GPT-Image vs Opus-4.1      -0.2232    0.0000
## X-squared87   Gemini-2.0-Flash-GPT-Image vs GPT-5         -0.3452    0.0000
## X-squared88   Sonnet-4 vs Opus-4.1                       -0.1154    0.0014
## X-squared89   Sonnet-4 vs GPT-5                          -0.2374    0.0000
## X-squared90   Opus-4.1 vs GPT-5                          -0.1220    0.0005
```

Visualization of Collapsed Data

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

print(collapsed_plot)
```



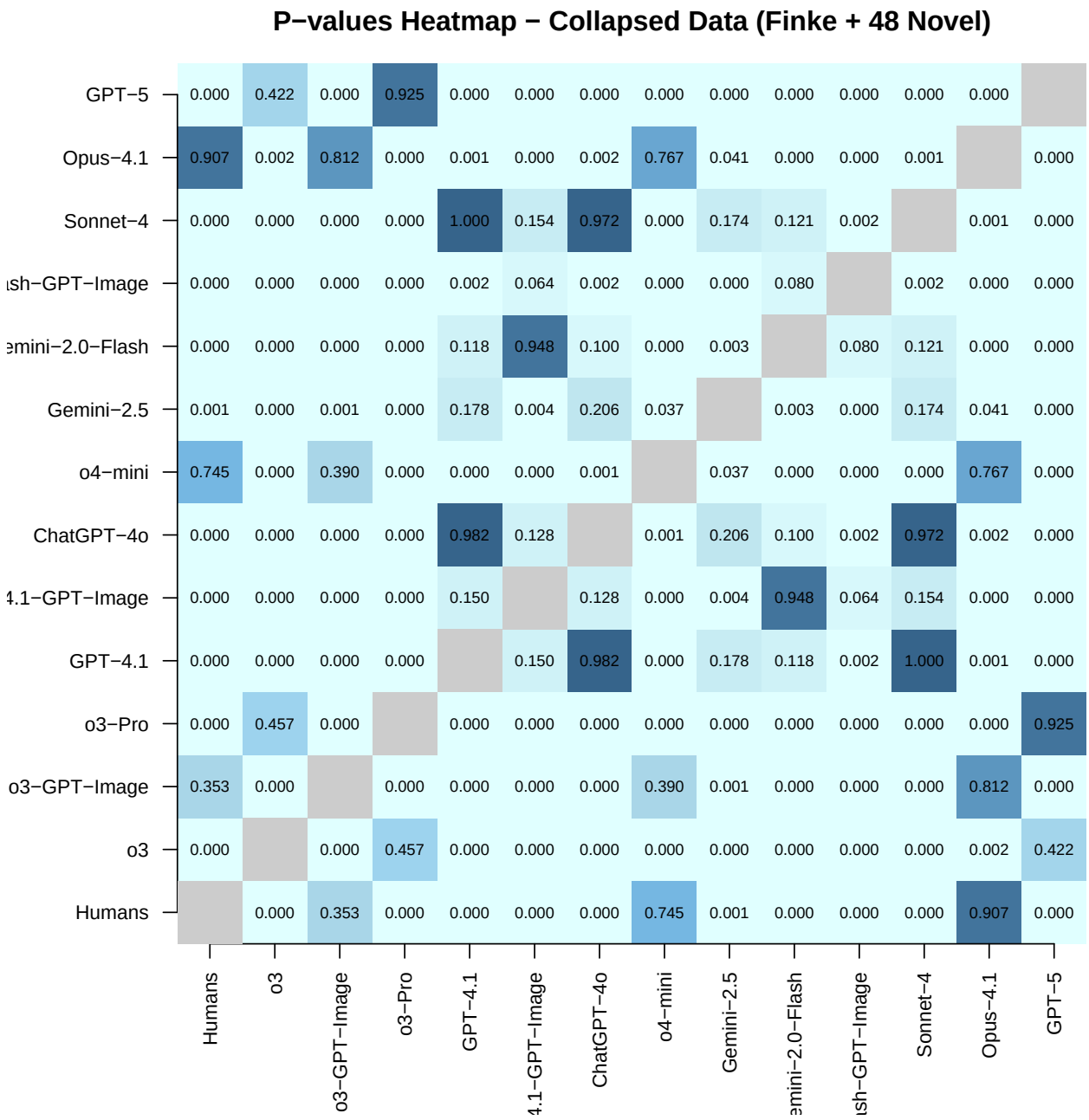

```r
# Create a comparison plot showing all three datasets

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

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

print(comparison_plot)
```



Heatmap for Collapsed Data

```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)
    }
  }
}
```



Reasoning Variation Analysis

Finke

```
# 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:
```

```
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:  0.624  vs  0.606
## Difference:  0.018
## Chi-squared:  0.151
## Degrees of freedom:  1
## P-value:  0.6979
## 95% CI: [ -0.061 ,  0.096 ]
## Significant:  NO
##
##  Humans vs o3-Medium
## ----------------------------------------
## Proportions:  0.624  vs  0.578
## Difference:  0.047
## Chi-squared:  0.353
## Degrees of freedom:  1
## P-value:  0.5526
## 95% CI: [ -0.089 ,  0.183 ]
## Significant:  NO
##
##  Humans vs o3-Low
## ----------------------------------------
## Proportions:  0.624  vs  0.627
## Difference:  -0.003
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.131 ,  0.125 ]
## Significant:  NO
##
##  Humans vs GPT-5-High
## ----------------------------------------
## Proportions:  0.624  vs  0.768
## Difference:  -0.144
## Chi-squared:  9.34
## Degrees of freedom:  1
## P-value:  0.002242
## 95% CI: [ -0.228 ,  -0.06 ]
## Significant:  YES (p < 0.05)
```

```
##
##  Humans vs GPT-5-Medium
## ---------------------------------------
## Proportions:  0.624  vs  0.633
## Difference:  -0.009
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  0.9939
## 95% CI: [ -0.142 ,  0.124 ]
## Significant:  NO
##
##  Humans vs GPT-5-Low
## ---------------------------------------
## Proportions:  0.624  vs  0.56
## Difference:  0.064
## Chi-squared:  0.757
## Degrees of freedom:  1
## P-value:  0.3843
## 95% CI: [ -0.072 ,  0.201 ]
## Significant:  NO
##
##  Humans vs GPT-5-Minimal
## ---------------------------------------
## Proportions:  0.624  vs  0.37
## Difference:  0.255
## Chi-squared:  14.741
## Degrees of freedom:  1
## P-value:  0.0001233
## 95% CI: [ 0.121 ,  0.388 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o4-mini-High
## ---------------------------------------
## Proportions:  0.624  vs  0.533
## Difference:  0.091
## Chi-squared:  3.519
## Degrees of freedom:  1
## P-value:  0.06067
## 95% CI: [ -0.006 ,  0.188 ]
## Significant:  NO
##
##  Humans vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.624  vs  0.461
## Difference:  0.164
## Chi-squared:  11.879
## Degrees of freedom:  1
## P-value:  0.0005677
## 95% CI: [ 0.067 ,  0.261 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.624  vs  0.568
```

```
## Difference:  0.057
## Chi-squared:  2.584
## Degrees of freedom:  1
## P-value:  0.1079
## 95% CI: [ -0.013 ,  0.126 ]
## Significant:  NO
##
##  Humans vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.624  vs  0.497
## Difference:  0.127
## Chi-squared:  3.431
## Degrees of freedom:  1
## P-value:  0.06398
## 95% CI: [ -0.01 ,  0.265 ]
## Significant:  NO
##
##  o3-High vs o3-Medium
## ---------------------------------------
## Proportions:  0.606  vs  0.578
## Difference:  0.029
## Chi-squared:  0.057
## Degrees of freedom:  1
## P-value:  0.8107
## 95% CI: [ -0.126 ,  0.184 ]
## Significant:  NO
##
##  o3-High vs o3-Low
## ---------------------------------------
## Proportions:  0.606  vs  0.627
## Difference:  -0.021
## Chi-squared:  0.018
## Degrees of freedom:  1
## P-value:  0.8936
## 95% CI: [ -0.174 ,  0.132 ]
## Significant:  NO
##
##  o3-High vs GPT-5-High
## ---------------------------------------
## Proportions:  0.606  vs  0.768
## Difference:  -0.162
## Chi-squared:  7.838
## Degrees of freedom:  1
## P-value:  0.005117
## 95% CI: [ -0.273 ,  -0.051 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs GPT-5-Medium
## ---------------------------------------
## Proportions:  0.606  vs  0.633
## Difference:  -0.027
## Chi-squared:  0.048
## Degrees of freedom:  1
## P-value:  0.8259
```

```
## 95% CI: [ -0.179 ,  0.125 ]
## Significant:  NO
##
##  o3-High vs GPT-5-Low
## ---------------------------------------
## Proportions:  0.606  vs  0.56
## Difference:  0.046
## Chi-squared:  0.231
## Degrees of freedom:  1
## P-value:  0.6311
## 95% CI: [ -0.109 ,  0.202 ]
## Significant:  NO
##
##  o3-High vs GPT-5-Minimal
## ---------------------------------------
## Proportions:  0.606  vs  0.37
## Difference:  0.237
## Chi-squared:  9.238
## Degrees of freedom:  1
## P-value:  0.00237
## 95% CI: [ 0.084 ,  0.389 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o4-mini-High
## ---------------------------------------
## Proportions:  0.606  vs  0.533
## Difference:  0.073
## Chi-squared:  1.287
## Degrees of freedom:  1
## P-value:  0.2566
## 95% CI: [ -0.048 ,  0.194 ]
## Significant:  NO
##
##  o3-High vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.606  vs  0.461
## Difference:  0.146
## Chi-squared:  5.602
## Degrees of freedom:  1
## P-value:  0.01794
## 95% CI: [ 0.025 ,  0.267 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.606  vs  0.568
## Difference:  0.039
## Chi-squared:  0.486
## Degrees of freedom:  1
## P-value:  0.4856
## 95% CI: [ -0.061 ,  0.139 ]
## Significant:  NO
##
##  o3-High vs o3-GPT-Image-Medium
```

```
## ----------------------------------------
## Proportions:  0.606  vs  0.497
## Difference:  0.109
## Chi-squared:  1.772
## Degrees of freedom:  1
## P-value:  0.1831
## 95% CI: [ -0.047 ,  0.265 ]
## Significant:  NO
##
##  o3-Medium vs o3-Low
## ----------------------------------------
## Proportions:  0.578  vs  0.627
## Difference:  -0.049
## Chi-squared:  0.135
## Degrees of freedom:  1
## P-value:  0.7137
## 95% CI: [ -0.241 ,  0.142 ]
## Significant:  NO
##
##  o3-Medium vs GPT-5-High
## ----------------------------------------
## Proportions:  0.578  vs  0.768
## Difference:  -0.191
## Chi-squared:  6.095
## Degrees of freedom:  1
## P-value:  0.01355
## 95% CI: [ -0.349 ,  -0.032 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.578  vs  0.633
## Difference:  -0.056
## Chi-squared:  0.191
## Degrees of freedom:  1
## P-value:  0.6618
## 95% CI: [ -0.247 ,  0.136 ]
## Significant:  NO
##
##  o3-Medium vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.578  vs  0.56
## Difference:  0.018
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  0.9914
## 95% CI: [ -0.176 ,  0.211 ]
## Significant:  NO
##
##  o3-Medium vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.578  vs  0.37
## Difference:  0.208
## Chi-squared:  4.407
```

```
## Degrees of freedom:  1
## P-value:  0.0358
## 95% CI: [ 0.017 ,  0.399 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs o4-mini-High
## ----------------------------------------
## Proportions:  0.578  vs  0.533
## Difference:  0.044
## Chi-squared:  0.164
## Degrees of freedom:  1
## P-value:  0.6854
## 95% CI: [ -0.122 ,  0.21 ]
## Significant:  NO
##
##  o3-Medium vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.578  vs  0.461
## Difference:  0.117
## Chi-squared:  1.752
## Degrees of freedom:  1
## P-value:  0.1856
## 95% CI: [ -0.049 ,  0.283 ]
## Significant:  NO
##
##  o3-Medium vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.578  vs  0.568
## Difference:  0.01
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.14 ,  0.16 ]
## Significant:  NO
##
##  o3-Medium vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.578  vs  0.497
## Difference:  0.08
## Chi-squared:  0.491
## Degrees of freedom:  1
## P-value:  0.4833
## 95% CI: [ -0.114 ,  0.275 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-High
## ----------------------------------------
## Proportions:  0.627  vs  0.768
## Difference:  -0.141
## Chi-squared:  3.291
## Degrees of freedom:  1
## P-value:  0.06965
## 95% CI: [ -0.297 ,  0.015 ]
## Significant:  NO
```

```
##
##  o3-Low vs GPT-5-Medium
## ---------------------------------------
## Proportions:  0.627  vs  0.633
## Difference:  -0.006
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.185 ,  0.173 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-Low
## ---------------------------------------
## Proportions:  0.627  vs  0.56
## Difference:  0.067
## Chi-squared:  0.316
## Degrees of freedom:  1
## P-value:  0.574
## 95% CI: [ -0.125 ,  0.259 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-Minimal
## ---------------------------------------
## Proportions:  0.627  vs  0.37
## Difference:  0.257
## Chi-squared:  6.959
## Degrees of freedom:  1
## P-value:  0.00834
## 95% CI: [ 0.068 ,  0.447 ]
## Significant:  YES (p < 0.05)
##
##  o3-Low vs o4-mini-High
## ---------------------------------------
## Proportions:  0.627  vs  0.533
## Difference:  0.094
## Chi-squared:  1.076
## Degrees of freedom:  1
## P-value:  0.2996
## 95% CI: [ -0.07 ,  0.258 ]
## Significant:  NO
##
##  o3-Low vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.627  vs  0.461
## Difference:  0.167
## Chi-squared:  3.803
## Degrees of freedom:  1
## P-value:  0.05115
## 95% CI: [ 0.003 ,  0.331 ]
## Significant:  NO
##
##  o3-Low vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.627  vs  0.568
```

```
## Difference:  0.06
## Chi-squared:  0.476
## Degrees of freedom:  1
## P-value:  0.4902
## 95% CI: [ -0.088 ,  0.207 ]
## Significant:  NO
##
##  o3-Low vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.627  vs  0.497
## Difference:  0.13
## Chi-squared:  1.563
## Degrees of freedom:  1
## P-value:  0.2112
## 95% CI: [ -0.063 ,  0.323 ]
## Significant:  NO
##
##  GPT-5-High vs GPT-5-Medium
## ---------------------------------------
## Proportions:  0.768  vs  0.633
## Difference:  0.135
## Chi-squared:  2.993
## Degrees of freedom:  1
## P-value:  0.08363
## 95% CI: [ -0.021 ,  0.291 ]
## Significant:  NO
##
##  GPT-5-High vs GPT-5-Low
## ---------------------------------------
## Proportions:  0.768  vs  0.56
## Difference:  0.208
## Chi-squared:  7.28
## Degrees of freedom:  1
## P-value:  0.006973
## 95% CI: [ 0.049 ,  0.367 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs GPT-5-Minimal
## ---------------------------------------
## Proportions:  0.768  vs  0.37
## Difference:  0.399
## Chi-squared:  25.74
## Degrees of freedom:  1
## P-value:  0.0000003906
## 95% CI: [ 0.243 ,  0.555 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o4-mini-High
## ---------------------------------------
## Proportions:  0.768  vs  0.533
## Difference:  0.235
## Chi-squared:  13.557
## Degrees of freedom:  1
## P-value:  0.0002314
```

```
## 95% CI: [ 0.11 ,  0.36 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.768  vs  0.461
## Difference:  0.308
## Chi-squared:  22.7
## Degrees of freedom:  1
## P-value:  0.000001894
## 95% CI: [ 0.183 ,  0.433 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.768  vs  0.568
## Difference:  0.201
## Chi-squared:  13.043
## Degrees of freedom:  1
## P-value:  0.0003044
## 95% CI: [ 0.096 ,  0.305 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.768  vs  0.497
## Difference:  0.271
## Chi-squared:  12.246
## Degrees of freedom:  1
## P-value:  0.0004663
## 95% CI: [ 0.111 ,  0.431 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs GPT-5-Low
## ---------------------------------------
## Proportions:  0.633  vs  0.56
## Difference:  0.073
## Chi-squared:  0.4
## Degrees of freedom:  1
## P-value:  0.5269
## 95% CI: [ -0.118 ,  0.265 ]
## Significant:  NO
##
##  GPT-5-Medium vs GPT-5-Minimal
## ---------------------------------------
## Proportions:  0.633  vs  0.37
## Difference:  0.264
## Chi-squared:  7.325
## Degrees of freedom:  1
## P-value:  0.0068
## 95% CI: [ 0.074 ,  0.453 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs o4-mini-High
```

```
## ----------------------------------------
## Proportions:  0.633  vs  0.533
## Difference:  0.1
## Chi-squared:  1.249
## Degrees of freedom:  1
## P-value:  0.2637
## 95% CI: [ -0.064 ,  0.264 ]
## Significant:  NO
##
##  GPT-5-Medium vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.633  vs  0.461
## Difference:  0.173
## Chi-squared:  4.119
## Degrees of freedom:  1
## P-value:  0.0424
## 95% CI: [ 0.009 ,  0.336 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.633  vs  0.568
## Difference:  0.066
## Chi-squared:  0.605
## Degrees of freedom:  1
## P-value:  0.4365
## 95% CI: [ -0.082 ,  0.213 ]
## Significant:  NO
##
##  GPT-5-Medium vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.633  vs  0.497
## Difference:  0.136
## Chi-squared:  1.744
## Degrees of freedom:  1
## P-value:  0.1867
## 95% CI: [ -0.056 ,  0.329 ]
## Significant:  NO
##
##  GPT-5-Low vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.56  vs  0.37
## Difference:  0.19
## Chi-squared:  3.64
## Degrees of freedom:  1
## P-value:  0.05642
## 95% CI: [ -0.001 ,  0.382 ]
## Significant:  NO
##
##  GPT-5-Low vs o4-mini-High
## ----------------------------------------
## Proportions:  0.56  vs  0.533
## Difference:  0.027
## Chi-squared:  0.033
```

```
## Degrees of freedom:  1
## P-value:  0.8566
## 95% CI: [ -0.14 ,  0.193 ]
## Significant:  NO
##
##  GPT-5-Low vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.56  vs  0.461
## Difference:  0.1
## Chi-squared:  1.212
## Degrees of freedom:  1
## P-value:  0.271
## 95% CI: [ -0.067 ,  0.266 ]
## Significant:  NO
##
##  GPT-5-Low vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.56  vs  0.568
## Difference:  -0.008
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.155 ,  0.14 ]
## Significant:  NO
##
##  GPT-5-Low vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.56  vs  0.497
## Difference:  0.063
## Chi-squared:  0.257
## Degrees of freedom:  1
## P-value:  0.6124
## 95% CI: [ -0.132 ,  0.258 ]
## Significant:  NO
##
##  GPT-5-Minimal vs o4-mini-High
## ---------------------------------------
## Proportions:  0.37  vs  0.533
## Difference:  -0.164
## Chi-squared:  3.662
## Degrees of freedom:  1
## P-value:  0.05565
## 95% CI: [ -0.327 ,  0 ]
## Significant:  NO
##
##  GPT-5-Minimal vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.37  vs  0.461
## Difference:  -0.091
## Chi-squared:  1.002
## Degrees of freedom:  1
## P-value:  0.3167
## 95% CI: [ -0.255 ,  0.073 ]
## Significant:  NO
```

```
##
##  GPT-5-Minimal vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.37  vs  0.568
## Difference:  -0.198
## Chi-squared:  6.77
## Degrees of freedom:  1
## P-value:  0.009271
## 95% CI: [ -0.346 ,  -0.05 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Minimal vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.37  vs  0.497
## Difference:  -0.128
## Chi-squared:  1.502
## Degrees of freedom:  1
## P-value:  0.2203
## 95% CI: [ -0.32 ,  0.065 ]
## Significant:  NO
##
##  o4-mini-High vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.533  vs  0.461
## Difference:  0.073
## Chi-squared:  0.997
## Degrees of freedom:  1
## P-value:  0.3181
## 95% CI: [ -0.062 ,  0.207 ]
## Significant:  NO
##
##  o4-mini-High vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.533  vs  0.568
## Difference:  -0.034
## Chi-squared:  0.254
## Degrees of freedom:  1
## P-value:  0.6144
## 95% CI: [ -0.15 ,  0.081 ]
## Significant:  NO
##
##  o4-mini-High vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.533  vs  0.497
## Difference:  0.036
## Chi-squared:  0.089
## Degrees of freedom:  1
## P-value:  0.7651
## 95% CI: [ -0.131 ,  0.203 ]
## Significant:  NO
##
##  o4-mini-Medium vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.461  vs  0.568
```

```
## Difference:  -0.107
## Chi-squared:  3.262
## Degrees of freedom:  1
## P-value:  0.07088
## 95% CI: [ -0.222 ,  0.008 ]
## Significant:  NO
##
##  o4-mini-Medium vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.461  vs  0.497
## Difference:  -0.037
## Chi-squared:  0.094
## Degrees of freedom:  1
## P-value:  0.7594
## 95% CI: [ -0.204 ,  0.131 ]
## Significant:  NO
##
##  o3-GPT-Image-High vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.568  vs  0.497
## Difference:  0.07
## Chi-squared:  0.697
## Degrees of freedom:  1
## P-value:  0.4037
## 95% CI: [ -0.081 ,  0.222 ]
## 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"))
```

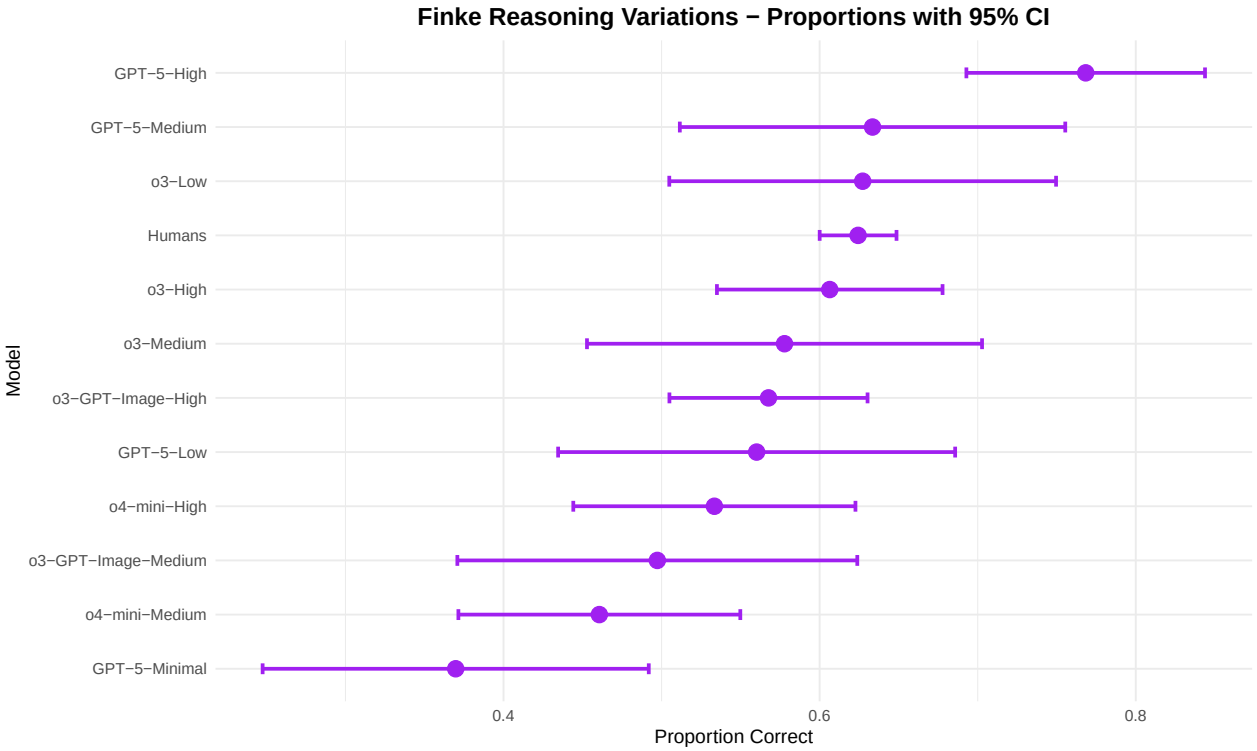
```
##
##
##                comparison                                diff   chi_squared   p_value  significant
## -------------  ----------------------------------------  -------  ------------  --------  -----------
## X-squared      Humans vs o3-High                          0.018     0.1506854    0.6979  FALSE
## X-squared1     Humans vs o3-Medium                        0.047     0.3526422    0.5526  FALSE
## X-squared2     Humans vs o3-Low                          -0.003     0.0000000    1.0000  FALSE
## X-squared3     Humans vs GPT-5-High                      -0.144     9.3403689    0.0022  TRUE
## X-squared4     Humans vs GPT-5-Medium                    -0.009     0.0000580    0.9939  FALSE
## X-squared5     Humans vs GPT-5-Low                        0.064     0.7569142    0.3843  FALSE
## X-squared6     Humans vs GPT-5-Minimal                    0.255    14.7407768    0.0001  TRUE
## X-squared7     Humans vs o4-mini-High                     0.091     3.5189352    0.0607  FALSE
## X-squared8     Humans vs o4-mini-Medium                   0.164    11.8789597    0.0006  TRUE
## X-squared9     Humans vs o3-GPT-Image-High                0.057     2.5840139    0.1079  FALSE
## X-squared10    Humans vs o3-GPT-Image-Medium              0.127     3.4310999    0.0640  FALSE
```

```
## X-squared11   o3-High vs o3-Medium                    0.029     0.0573855    0.8107  FALSE
## X-squared12   o3-High vs o3-Low                      -0.021     0.0179059    0.8936  FALSE
## X-squared13   o3-High vs GPT-5-High                  -0.162     7.8377188    0.0051  TRUE
## X-squared14   o3-High vs GPT-5-Medium                -0.027     0.0483992    0.8259  FALSE
## X-squared15   o3-High vs GPT-5-Low                    0.046     0.2305458    0.6311  FALSE
## X-squared16   o3-High vs GPT-5-Minimal                0.237     9.2383658    0.0024  TRUE
## X-squared17   o3-High vs o4-mini-High                 0.073     1.2867842    0.2566  FALSE
## X-squared18   o3-High vs o4-mini-Medium               0.146     5.6018071    0.0179  TRUE
## X-squared19   o3-High vs o3-GPT-Image-High           0.039     0.4863142    0.4856  FALSE
## X-squared20   o3-High vs o3-GPT-Image-Medium         0.109     1.7722539    0.1831  FALSE
## X-squared21   o3-Medium vs o3-Low                   -0.049     0.1345817    0.7137  FALSE
## X-squared22   o3-Medium vs GPT-5-High               -0.191     6.0953896    0.0136  TRUE
## X-squared23   o3-Medium vs GPT-5-Medium             -0.056     0.1913180    0.6618  FALSE
## X-squared24   o3-Medium vs GPT-5-Low                0.018     0.0001156    0.9914  FALSE
## X-squared25   o3-Medium vs GPT-5-Minimal            0.208     4.4067905    0.0358  TRUE
## X-squared26   o3-Medium vs o4-mini-High             0.044     0.1640870    0.6854  FALSE
## X-squared27   o3-Medium vs o4-mini-Medium           0.117     1.7523534    0.1856  FALSE
## X-squared28   o3-Medium vs o3-GPT-Image-High       0.010     0.0000000    1.0000  FALSE
## X-squared29   o3-Medium vs o3-GPT-Image-Medium     0.080     0.4913684    0.4833  FALSE
## X-squared30   o3-Low vs GPT-5-High                 -0.141     3.2912474    0.0697  FALSE
## X-squared31   o3-Low vs GPT-5-Medium               -0.006     0.0000000    1.0000  FALSE
## X-squared32   o3-Low vs GPT-5-Low                  0.067     0.3161179    0.5740  FALSE
## X-squared33   o3-Low vs GPT-5-Minimal              0.257     6.9590007    0.0083  TRUE
## X-squared34   o3-Low vs o4-mini-High               0.094     1.0760609    0.2996  FALSE
## X-squared35   o3-Low vs o4-mini-Medium             0.167     3.8033242    0.0512  FALSE
## X-squared36   o3-Low vs o3-GPT-Image-High         0.060     0.4761887    0.4902  FALSE
## X-squared37   o3-Low vs o3-GPT-Image-Medium       0.130     1.5634180    0.2112  FALSE
## X-squared38   GPT-5-High vs GPT-5-Medium           0.135     2.9929288    0.0836  FALSE
## X-squared39   GPT-5-High vs GPT-5-Low              0.208     7.2798106    0.0070  TRUE
## X-squared40   GPT-5-High vs GPT-5-Minimal          0.399    25.7401700    0.0000  TRUE
## X-squared41   GPT-5-High vs o4-mini-High           0.235    13.5574456    0.0002  TRUE
## X-squared42   GPT-5-High vs o4-mini-Medium         0.308    22.7000752    0.0000  TRUE
## X-squared43   GPT-5-High vs o3-GPT-Image-High     0.201    13.0430968    0.0003  TRUE
## X-squared44   GPT-5-High vs o3-GPT-Image-Medium   0.271    12.2460123    0.0005  TRUE
## X-squared45   GPT-5-Medium vs GPT-5-Low            0.073     0.4003402    0.5269  FALSE
## X-squared46   GPT-5-Medium vs GPT-5-Minimal        0.264     7.3249200    0.0068  TRUE
## X-squared47   GPT-5-Medium vs o4-mini-High         0.100     1.2492152    0.2637  FALSE
## X-squared48   GPT-5-Medium vs o4-mini-Medium       0.173     4.1192794    0.0424  TRUE
## X-squared49   GPT-5-Medium vs o3-GPT-Image-High   0.066     0.6054465    0.4365  FALSE
## X-squared50   GPT-5-Medium vs o3-GPT-Image-Medium 0.136     1.7435618    0.1867  FALSE
## X-squared51   GPT-5-Low vs GPT-5-Minimal           0.190     3.6397557    0.0564  FALSE
## X-squared52   GPT-5-Low vs o4-mini-High            0.027     0.0326606    0.8566  FALSE
## X-squared53   GPT-5-Low vs o4-mini-Medium          0.100     1.2116132    0.2710  FALSE
## X-squared54   GPT-5-Low vs o3-GPT-Image-High      -0.008     0.0000000    1.0000  FALSE
## X-squared55   GPT-5-Low vs o3-GPT-Image-Medium     0.063     0.2566541    0.6124  FALSE
## X-squared56   GPT-5-Minimal vs o4-mini-High       -0.164     3.6624836    0.0557  FALSE
## X-squared57   GPT-5-Minimal vs o4-mini-Medium     -0.091     1.0024851    0.3167  FALSE
## X-squared58   GPT-5-Minimal vs o3-GPT-Image-High  -0.198     6.7698922    0.0093  TRUE
## X-squared59   GPT-5-Minimal vs o3-GPT-Image-Medium -0.128    1.5022225    0.2203  FALSE
## X-squared60   o4-mini-High vs o4-mini-Medium       0.073     0.9967763    0.3181  FALSE
## X-squared61   o4-mini-High vs o3-GPT-Image-High   -0.034     0.2538412    0.6144  FALSE
## X-squared62   o4-mini-High vs o3-GPT-Image-Medium  0.036     0.0892857    0.7651  FALSE
## X-squared63   o4-mini-Medium vs o3-GPT-Image-High -0.107     3.2624114    0.0709  FALSE
## X-squared64   o4-mini-Medium vs o3-GPT-Image-Medium -0.037   0.0937977    0.7594  FALSE
```

```
## X-squared65   o3-GPT-Image-High vs o3-GPT-Image-Medium     0.070     0.6973783     0.4037  FALSE
```

Visualization of Finke Reasoning Variations

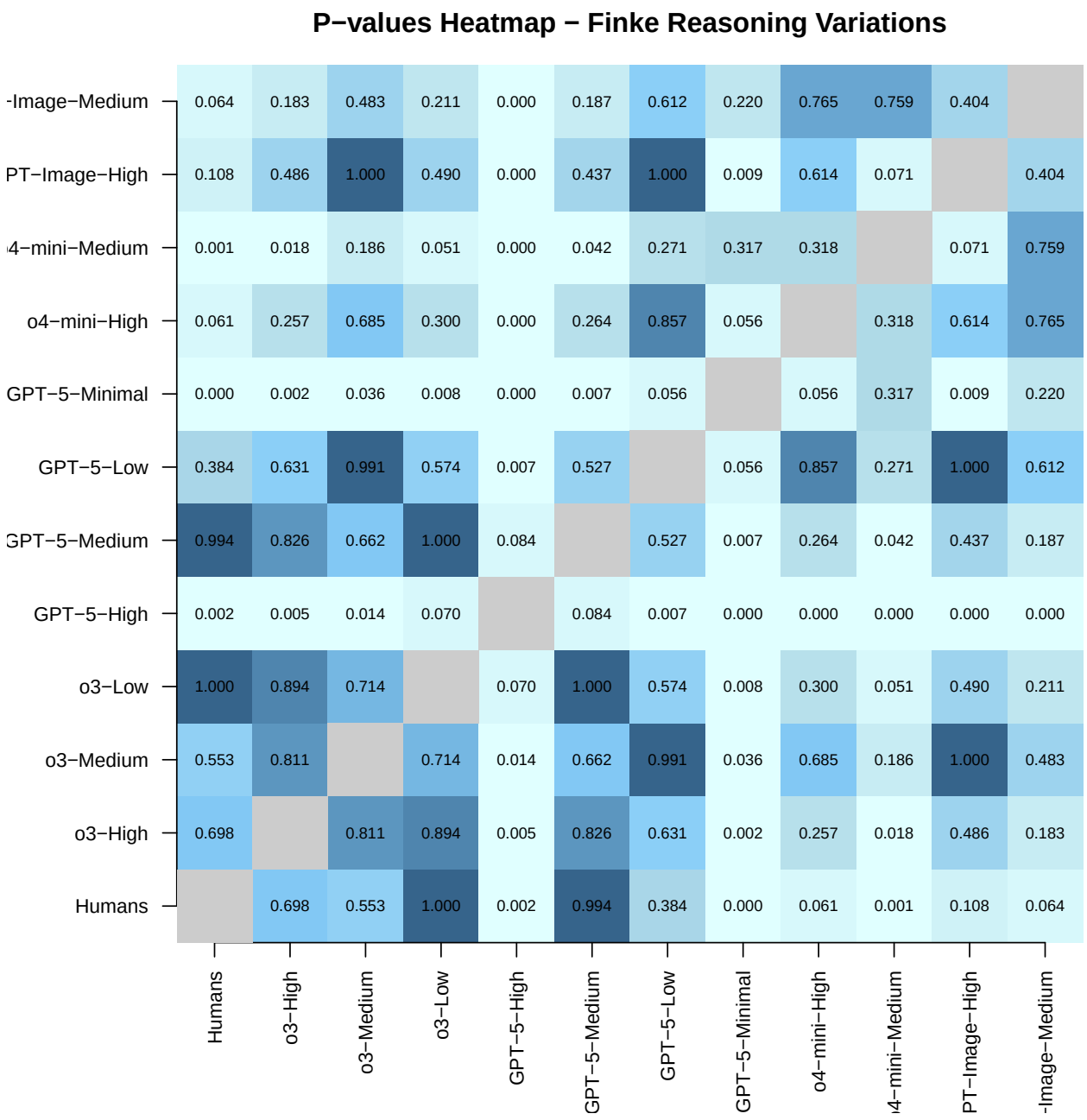
```r
# Plot proportions with confidence intervals for Finke reasoning variations
finke_reasoning_plot <- ggplot(finke_reasoning_data, aes(x = reorder(model, proportion), y = proportion
  geom_point(size = 4, color = "purple") +
  geom_errorbar(aes(ymin = proportion - 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    ymax = proportion + 1.96 * sqrt(proportion * (1 - proportion) / max_score)),
                width = 0.2, size = 1, color = "purple") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Finke Reasoning Variations - Proportions with 95% CI",
       x = "Model",
       y = "Proportion Correct") +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))
print(finke_reasoning_plot)
```



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

```
# 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:
```

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

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

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

```
##   Total comparisons: 66
```

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

```
##   Significant differences: 18
```

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

```
##   Percentage significant: 27.3 %
```

```
# 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                -0.1440    0.0022
## X-squared6     Humans vs GPT-5-Minimal              0.2546    0.0001
## X-squared8     Humans vs o4-mini-Medium             0.1637    0.0006
## X-squared13    o3-High vs GPT-5-High               -0.1619    0.0051
## X-squared16    o3-High vs GPT-5-Minimal             0.2366    0.0024
## X-squared18    o3-High vs o4-mini-Medium            0.1458    0.0179
## X-squared22    o3-Medium vs GPT-5-High             -0.1906    0.0136
## X-squared25    o3-Medium vs GPT-5-Minimal           0.2080    0.0358
## X-squared33    o3-Low vs GPT-5-Minimal              0.2575    0.0083
## X-squared39    GPT-5-High vs GPT-5-Low              0.2082    0.0070
## X-squared40    GPT-5-High vs GPT-5-Minimal          0.3986    0.0000
## X-squared41    GPT-5-High vs o4-mini-High           0.2349    0.0002
## X-squared42    GPT-5-High vs o4-mini-Medium         0.3077    0.0000
## X-squared43    GPT-5-High vs o3-GPT-Image-High      0.2007    0.0003
## X-squared44    GPT-5-High vs o3-GPT-Image-Medium    0.2710    0.0005
## X-squared46    GPT-5-Medium vs GPT-5-Minimal        0.2637    0.0068
## X-squared48    GPT-5-Medium vs o4-mini-Medium       0.1728    0.0424
## X-squared58    GPT-5-Minimal vs o3-GPT-Image-High  -0.1979    0.0093
```

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")
```

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

```r
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:  0.52  vs  0.657
## Difference:  -0.137
## Chi-squared:  47.906
## Degrees of freedom:  1
## P-value:  0.000000000004471
## 95% CI: [ -0.175 ,  -0.099 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o3-Medium
## ----------------------------------------
## Proportions:  0.52  vs  0.57
## Difference:  -0.051
## Chi-squared:  2.168
## Degrees of freedom:  1
## P-value:  0.1409
## 95% CI: [ -0.117 ,  0.015 ]
## Significant:  NO
##
##  Humans vs o3-Low
## ----------------------------------------
## Proportions:  0.52  vs  0.528
## Difference:  -0.008
## Chi-squared:  0.033
## Degrees of freedom:  1
## P-value:  0.8553
## 95% CI: [ -0.075 ,  0.058 ]
## Significant:  NO
##
##  Humans vs GPT-5-High
## ----------------------------------------
## Proportions:  0.52  vs  0.643
## Difference:  -0.123
## Chi-squared:  26.644
## Degrees of freedom:  1
## P-value:  0.0000002446
## 95% CI: [ -0.169 ,  -0.078 ]
## Significant:  YES (p < 0.05)
```

```
##
##  Humans vs GPT-5-Medium
## ---------------------------------------
## Proportions:  0.52  vs  0.584
## Difference:  -0.064
## Chi-squared:  3.573
## Degrees of freedom:  1
## P-value:  0.05872
## 95% CI: [ -0.13 ,  0.001 ]
## Significant:  NO
##
##  Humans vs GPT-5-Low
## ---------------------------------------
## Proportions:  0.52  vs  0.497
## Difference:  0.023
## Chi-squared:  0.383
## Degrees of freedom:  1
## P-value:  0.5358
## 95% CI: [ -0.044 ,  0.089 ]
## Significant:  NO
##
##  Humans vs GPT-5-Minimal
## ---------------------------------------
## Proportions:  0.52  vs  0.418
## Difference:  0.102
## Chi-squared:  9.17
## Degrees of freedom:  1
## P-value:  0.00246
## 95% CI: [ 0.036 ,  0.168 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o4-mini-High
## ---------------------------------------
## Proportions:  0.52  vs  0.533
## Difference:  -0.013
## Chi-squared:  0.271
## Degrees of freedom:  1
## P-value:  0.6023
## 95% CI: [ -0.061 ,  0.034 ]
## Significant:  NO
##
##  Humans vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.52  vs  0.496
## Difference:  0.024
## Chi-squared:  0.923
## Degrees of freedom:  1
## P-value:  0.3367
## 95% CI: [ -0.024 ,  0.072 ]
## Significant:  NO
##
##  Humans vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.52  vs  0.553
```

```
## Difference:  -0.033
## Chi-squared:  3.527
## Degrees of freedom:  1
## P-value:  0.06039
## 95% CI: [ -0.068 ,  0.001 ]
## Significant:  NO
##
##  Humans vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.52  vs  0.57
## Difference:  -0.05
## Chi-squared:  2.119
## Degrees of freedom:  1
## P-value:  0.1454
## 95% CI: [ -0.116 ,  0.016 ]
## Significant:  NO
##
##  o3-High vs o3-Medium
## ---------------------------------------
## Proportions:  0.657  vs  0.57
## Difference:  0.086
## Chi-squared:  5.431
## Degrees of freedom:  1
## P-value:  0.01979
## 95% CI: [ 0.012 ,  0.161 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o3-Low
## ---------------------------------------
## Proportions:  0.657  vs  0.528
## Difference:  0.129
## Chi-squared:  12.194
## Degrees of freedom:  1
## P-value:  0.0004794
## 95% CI: [ 0.054 ,  0.204 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs GPT-5-High
## ---------------------------------------
## Proportions:  0.657  vs  0.643
## Difference:  0.014
## Chi-squared:  0.179
## Degrees of freedom:  1
## P-value:  0.6722
## 95% CI: [ -0.043 ,  0.07 ]
## Significant:  NO
##
##  o3-High vs GPT-5-Medium
## ---------------------------------------
## Proportions:  0.657  vs  0.584
## Difference:  0.073
## Chi-squared:  3.808
## Degrees of freedom:  1
## P-value:  0.05101
```

```
## 95% CI: [ -0.001 ,  0.147 ]
## Significant:  NO
##
##  o3-High vs GPT-5-Low
## ---------------------------------------
## Proportions:  0.657  vs  0.497
## Difference:  0.16
## Chi-squared:  18.71
## Degrees of freedom:  1
## P-value:  0.00001522
## 95% CI: [ 0.085 ,  0.234 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs GPT-5-Minimal
## ---------------------------------------
## Proportions:  0.657  vs  0.418
## Difference:  0.239
## Chi-squared:  41.67
## Degrees of freedom:  1
## P-value:  0.000000000108
## 95% CI: [ 0.165 ,  0.313 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o4-mini-High
## ---------------------------------------
## Proportions:  0.657  vs  0.533
## Difference:  0.124
## Chi-squared:  17.907
## Degrees of freedom:  1
## P-value:  0.00002319
## 95% CI: [ 0.065 ,  0.182 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.657  vs  0.496
## Difference:  0.161
## Chi-squared:  30.21
## Degrees of freedom:  1
## P-value:  0.00000003878
## 95% CI: [ 0.103 ,  0.219 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.657  vs  0.553
## Difference:  0.104
## Chi-squared:  17.991
## Degrees of freedom:  1
## P-value:  0.0000222
## 95% CI: [ 0.056 ,  0.152 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o3-GPT-Image-Medium
```

```
## ----------------------------------------
## Proportions:  0.657  vs  0.57
## Difference:  0.087
## Chi-squared:  5.501
## Degrees of freedom:  1
## P-value:  0.01901
## 95% CI: [ 0.013 ,  0.161 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs o3-Low
## ----------------------------------------
## Proportions:  0.57  vs  0.528
## Difference:  0.042
## Chi-squared:  0.709
## Degrees of freedom:  1
## P-value:  0.3996
## 95% CI: [ -0.051 ,  0.136 ]
## Significant:  NO
##
##  o3-Medium vs GPT-5-High
## ----------------------------------------
## Proportions:  0.57  vs  0.643
## Difference:  -0.073
## Chi-squared:  3.291
## Degrees of freedom:  1
## P-value:  0.06968
## 95% CI: [ -0.152 ,  0.006 ]
## Significant:  NO
##
##  o3-Medium vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.57  vs  0.584
## Difference:  -0.014
## Chi-squared:  0.045
## Degrees of freedom:  1
## P-value:  0.8319
## 95% CI: [ -0.106 ,  0.079 ]
## Significant:  NO
##
##  o3-Medium vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.57  vs  0.497
## Difference:  0.073
## Chi-squared:  2.293
## Degrees of freedom:  1
## P-value:  0.13
## 95% CI: [ -0.02 ,  0.166 ]
## Significant:  NO
##
##  o3-Medium vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.57  vs  0.418
## Difference:  0.152
## Chi-squared:  10.547
```

```
## Degrees of freedom:  1
## P-value:  0.001164
## 95% CI: [ 0.06 ,  0.245 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs o4-mini-High
## ----------------------------------------
## Proportions:  0.57  vs  0.533
## Difference:  0.037
## Chi-squared:  0.746
## Degrees of freedom:  1
## P-value:  0.3879
## 95% CI: [ -0.043 ,  0.117 ]
## Significant:  NO
##
##  o3-Medium vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.57  vs  0.496
## Difference:  0.074
## Chi-squared:  3.265
## Degrees of freedom:  1
## P-value:  0.07076
## 95% CI: [ -0.006 ,  0.155 ]
## Significant:  NO
##
##  o3-Medium vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.57  vs  0.553
## Difference:  0.017
## Chi-squared:  0.17
## Degrees of freedom:  1
## P-value:  0.6804
## 95% CI: [ -0.055 ,  0.09 ]
## Significant:  NO
##
##  o3-Medium vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.57  vs  0.57
## Difference:  0.001
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.089 ,  0.09 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-High
## ----------------------------------------
## Proportions:  0.528  vs  0.643
## Difference:  -0.115
## Chi-squared:  8.407
## Degrees of freedom:  1
## P-value:  0.003737
## 95% CI: [ -0.195 ,  -0.036 ]
## Significant:  YES (p < 0.05)
```

```
##
##  o3-Low vs GPT-5-Medium
## ---------------------------------------
## Proportions:  0.528  vs  0.584
## Difference:  -0.056
## Chi-squared:  1.314
## Degrees of freedom:  1
## P-value:  0.2516
## 95% CI: [ -0.149 ,  0.037 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-Low
## ---------------------------------------
## Proportions:  0.528  vs  0.497
## Difference:  0.031
## Chi-squared:  0.338
## Degrees of freedom:  1
## P-value:  0.5609
## 95% CI: [ -0.063 ,  0.124 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-Minimal
## ---------------------------------------
## Proportions:  0.528  vs  0.418
## Difference:  0.11
## Chi-squared:  5.389
## Degrees of freedom:  1
## P-value:  0.02027
## 95% CI: [ 0.017 ,  0.203 ]
## Significant:  YES (p < 0.05)
##
##  o3-Low vs o4-mini-High
## ---------------------------------------
## Proportions:  0.528  vs  0.533
## Difference:  -0.005
## Chi-squared:  0.003
## Degrees of freedom:  1
## P-value:  0.9558
## 95% CI: [ -0.086 ,  0.075 ]
## Significant:  NO
##
##  o3-Low vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.528  vs  0.496
## Difference:  0.032
## Chi-squared:  0.536
## Degrees of freedom:  1
## P-value:  0.464
## 95% CI: [ -0.048 ,  0.113 ]
## Significant:  NO
##
##  o3-Low vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.528  vs  0.553
```

```
## Difference:  -0.025
## Chi-squared:  0.391
## Degrees of freedom:  1
## P-value:  0.532
## 95% CI: [ -0.098 ,  0.048 ]
## Significant:  NO
##
##  o3-Low vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.528  vs  0.57
## Difference:  -0.042
## Chi-squared:  0.689
## Degrees of freedom:  1
## P-value:  0.4064
## 95% CI: [ -0.135 ,  0.051 ]
## Significant:  NO
##
##  GPT-5-High vs GPT-5-Medium
## ---------------------------------------
## Proportions:  0.643  vs  0.584
## Difference:  0.059
## Chi-squared:  2.13
## Degrees of freedom:  1
## P-value:  0.1444
## 95% CI: [ -0.02 ,  0.138 ]
## Significant:  NO
##
##  GPT-5-High vs GPT-5-Low
## ---------------------------------------
## Proportions:  0.643  vs  0.497
## Difference:  0.146
## Chi-squared:  13.529
## Degrees of freedom:  1
## P-value:  0.0002349
## 95% CI: [ 0.066 ,  0.225 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs GPT-5-Minimal
## ---------------------------------------
## Proportions:  0.643  vs  0.418
## Difference:  0.225
## Chi-squared:  32.148
## Degrees of freedom:  1
## P-value:  0.00000001429
## 95% CI: [ 0.146 ,  0.304 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o4-mini-High
## ---------------------------------------
## Proportions:  0.643  vs  0.533
## Difference:  0.11
## Chi-squared:  11.515
## Degrees of freedom:  1
## P-value:  0.0006902
```

```
## 95% CI: [ 0.046 ,  0.174 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.643  vs  0.496
## Difference:  0.147
## Chi-squared:  20.634
## Degrees of freedom:  1
## P-value:  0.00000556
## 95% CI: [ 0.083 ,  0.211 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.643  vs  0.553
## Difference:  0.09
## Chi-squared:  10.329
## Degrees of freedom:  1
## P-value:  0.001309
## 95% CI: [ 0.035 ,  0.145 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.643  vs  0.57
## Difference:  0.073
## Chi-squared:  3.342
## Degrees of freedom:  1
## P-value:  0.06754
## 95% CI: [ -0.006 ,  0.152 ]
## Significant:  NO
##
##  GPT-5-Medium vs GPT-5-Low
## ---------------------------------------
## Proportions:  0.584  vs  0.497
## Difference:  0.087
## Chi-squared:  3.304
## Degrees of freedom:  1
## P-value:  0.06909
## 95% CI: [ -0.006 ,  0.18 ]
## Significant:  NO
##
##  GPT-5-Medium vs GPT-5-Minimal
## ---------------------------------------
## Proportions:  0.584  vs  0.418
## Difference:  0.166
## Chi-squared:  12.592
## Degrees of freedom:  1
## P-value:  0.0003875
## 95% CI: [ 0.074 ,  0.259 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs o4-mini-High
```

```
## ----------------------------------------
## Proportions:  0.584  vs  0.533
## Difference:  0.051
## Chi-squared:  1.473
## Degrees of freedom:  1
## P-value:  0.2249
## 95% CI: [ -0.029 ,  0.131 ]
## Significant:  NO
##
##  GPT-5-Medium vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.584  vs  0.496
## Difference:  0.088
## Chi-squared:  4.648
## Degrees of freedom:  1
## P-value:  0.0311
## 95% CI: [ 0.008 ,  0.168 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.584  vs  0.553
## Difference:  0.031
## Chi-squared:  0.633
## Degrees of freedom:  1
## P-value:  0.4263
## 95% CI: [ -0.041 ,  0.104 ]
## Significant:  NO
##
##  GPT-5-Medium vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.584  vs  0.57
## Difference:  0.014
## Chi-squared:  0.05
## Degrees of freedom:  1
## P-value:  0.8225
## 95% CI: [ -0.078 ,  0.107 ]
## Significant:  NO
##
##  GPT-5-Low vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.497  vs  0.418
## Difference:  0.079
## Chi-squared:  2.727
## Degrees of freedom:  1
## P-value:  0.09867
## 95% CI: [ -0.014 ,  0.172 ]
## Significant:  NO
##
##  GPT-5-Low vs o4-mini-High
## ----------------------------------------
## Proportions:  0.497  vs  0.533
## Difference:  -0.036
## Chi-squared:  0.693
```

```
## Degrees of freedom:  1
## P-value:  0.405
## 95% CI: [ -0.117 ,  0.045 ]
## Significant:  NO
##
##  GPT-5-Low vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.497  vs  0.496
## Difference:  0.001
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.077 ,  0.08 ]
## Significant:  NO
##
##  GPT-5-Low vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.497  vs  0.553
## Difference:  -0.056
## Chi-squared:  2.185
## Degrees of freedom:  1
## P-value:  0.1394
## 95% CI: [ -0.129 ,  0.017 ]
## Significant:  NO
##
##  GPT-5-Low vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.497  vs  0.57
## Difference:  -0.073
## Chi-squared:  2.257
## Degrees of freedom:  1
## P-value:  0.133
## 95% CI: [ -0.166 ,  0.021 ]
## Significant:  NO
##
##  GPT-5-Minimal vs o4-mini-High
## ---------------------------------------
## Proportions:  0.418  vs  0.533
## Difference:  -0.115
## Chi-squared:  8.051
## Degrees of freedom:  1
## P-value:  0.004549
## 95% CI: [ -0.195 ,  -0.035 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Minimal vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.418  vs  0.496
## Difference:  -0.078
## Chi-squared:  3.591
## Degrees of freedom:  1
## P-value:  0.05808
## 95% CI: [ -0.158 ,  0.002 ]
## Significant:  NO
```

```
##
##  GPT-5-Minimal vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.418  vs  0.553
## Difference:  -0.135
## Chi-squared:  13.502
## Degrees of freedom:  1
## P-value:  0.0002383
## 95% CI: [ -0.207 ,  -0.063 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Minimal vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.418  vs  0.57
## Difference:  -0.152
## Chi-squared:  10.47
## Degrees of freedom:  1
## P-value:  0.001213
## 95% CI: [ -0.244 ,  -0.059 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-High vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.533  vs  0.496
## Difference:  0.037
## Chi-squared:  1.197
## Degrees of freedom:  1
## P-value:  0.2739
## 95% CI: [ -0.028 ,  0.103 ]
## Significant:  NO
##
##  o4-mini-High vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.533  vs  0.553
## Difference:  -0.02
## Chi-squared:  0.427
## Degrees of freedom:  1
## P-value:  0.5136
## 95% CI: [ -0.076 ,  0.036 ]
## Significant:  NO
##
##  o4-mini-High vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.533  vs  0.57
## Difference:  -0.037
## Chi-squared:  0.722
## Degrees of freedom:  1
## P-value:  0.3956
## 95% CI: [ -0.117 ,  0.043 ]
## Significant:  NO
##
##  o4-mini-Medium vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.496  vs  0.553
```

```
## Difference:  -0.057
## Chi-squared:  3.969
## Degrees of freedom:  1
## P-value:  0.04634
## 95% CI: [ -0.113 ,  -0.001 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-Medium vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.496  vs  0.57
## Difference:  -0.074
## Chi-squared:  3.216
## Degrees of freedom:  1
## P-value:  0.07294
## 95% CI: [ -0.154 ,  0.006 ]
## Significant:  NO
##
##  o3-GPT-Image-High vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.553  vs  0.57
## Difference:  -0.017
## Chi-squared:  0.157
## Degrees of freedom:  1
## P-value:  0.6916
## 95% CI: [ -0.09 ,  0.056 ]
## 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),
         p_value = round(p_value, 4))
cat("\n\nSummary Table - 48 Novel Reasoning Variations:\n")
```

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

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

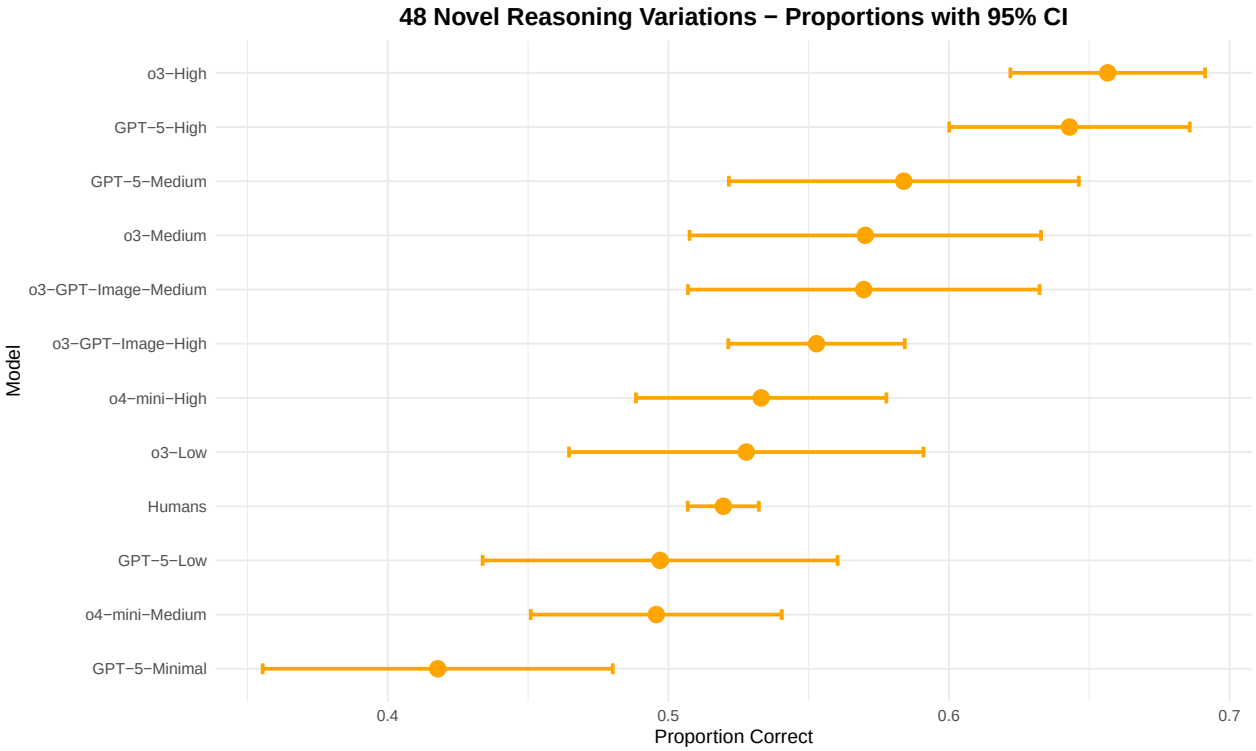
```
##
##
##                comparison                                diff   chi_squared   p_value  significant
## -------------  ----------------------------------------  -------  ------------  --------  -----------
## X-squared      Humans vs o3-High                        -0.137    47.9063736    0.0000  TRUE
## X-squared1     Humans vs o3-Medium                      -0.051     2.1680212    0.1409  FALSE
## X-squared2     Humans vs o3-Low                         -0.008     0.0332551    0.8553  FALSE
## X-squared3     Humans vs GPT-5-High                     -0.123    26.6438423    0.0000  TRUE
## X-squared4     Humans vs GPT-5-Medium                   -0.064     3.5732443    0.0587  FALSE
## X-squared5     Humans vs GPT-5-Low                       0.023     0.3833560    0.5358  FALSE
## X-squared6     Humans vs GPT-5-Minimal                   0.102     9.1700343    0.0025  TRUE
## X-squared7     Humans vs o4-mini-High                   -0.013     0.2714854    0.6023  FALSE
## X-squared8     Humans vs o4-mini-Medium                  0.024     0.9231232    0.3367  FALSE
## X-squared9     Humans vs o3-GPT-Image-High              -0.033     3.5265237    0.0604  FALSE
## X-squared10    Humans vs o3-GPT-Image-Medium            -0.050     2.1194305    0.1454  FALSE
```

```
## X-squared11   o3-High vs o3-Medium                    0.086     5.4305770    0.0198  TRUE
## X-squared12   o3-High vs o3-Low                       0.129    12.1940369    0.0005  TRUE
## X-squared13   o3-High vs GPT-5-High                   0.014     0.1790592    0.6722  FALSE
## X-squared14   o3-High vs GPT-5-Medium                 0.073     3.8080201    0.0510  FALSE
## X-squared15   o3-High vs GPT-5-Low                    0.160    18.7095412    0.0000  TRUE
## X-squared16   o3-High vs GPT-5-Minimal                0.239    41.6703542    0.0000  TRUE
## X-squared17   o3-High vs o4-mini-High                 0.124    17.9073744    0.0000  TRUE
## X-squared18   o3-High vs o4-mini-Medium               0.161    30.2095792    0.0000  TRUE
## X-squared19   o3-High vs o3-GPT-Image-High           0.104    17.9907647    0.0000  TRUE
## X-squared20   o3-High vs o3-GPT-Image-Medium         0.087     5.5008033    0.0190  TRUE
## X-squared21   o3-Medium vs o3-Low                    0.042     0.7094029    0.3996  FALSE
## X-squared22   o3-Medium vs GPT-5-High               -0.073     3.2906560    0.0697  FALSE
## X-squared23   o3-Medium vs GPT-5-Medium             -0.014     0.0450623    0.8319  FALSE
## X-squared24   o3-Medium vs GPT-5-Low                0.073     2.2928881    0.1300  FALSE
## X-squared25   o3-Medium vs GPT-5-Minimal            0.152    10.5473242    0.0012  TRUE
## X-squared26   o3-Medium vs o4-mini-High             0.037     0.7455878    0.3879  FALSE
## X-squared27   o3-Medium vs o4-mini-Medium           0.074     3.2653665    0.0708  FALSE
## X-squared28   o3-Medium vs o3-GPT-Image-High       0.017     0.1696483    0.6804  FALSE
## X-squared29   o3-Medium vs o3-GPT-Image-Medium     0.001     0.0000000    1.0000  FALSE
## X-squared30   o3-Low vs GPT-5-High                 -0.115     8.4074352    0.0037  TRUE
## X-squared31   o3-Low vs GPT-5-Medium               -0.056     1.3142630    0.2516  FALSE
## X-squared32   o3-Low vs GPT-5-Low                  0.031     0.3381216    0.5609  FALSE
## X-squared33   o3-Low vs GPT-5-Minimal              0.110     5.3887825    0.0203  TRUE
## X-squared34   o3-Low vs o4-mini-High               -0.005     0.0030676    0.9558  FALSE
## X-squared35   o3-Low vs o4-mini-Medium             0.032     0.5362303    0.4640  FALSE
## X-squared36   o3-Low vs o3-GPT-Image-High         -0.025     0.3906149    0.5320  FALSE
## X-squared37   o3-Low vs o3-GPT-Image-Medium       -0.042     0.6892385    0.4064  FALSE
## X-squared38   GPT-5-High vs GPT-5-Medium           0.059     2.1304888    0.1444  FALSE
## X-squared39   GPT-5-High vs GPT-5-Low              0.146    13.5286768    0.0002  TRUE
## X-squared40   GPT-5-High vs GPT-5-Minimal          0.225    32.1478468    0.0000  TRUE
## X-squared41   GPT-5-High vs o4-mini-High           0.110    11.5154242    0.0007  TRUE
## X-squared42   GPT-5-High vs o4-mini-Medium         0.147    20.6338825    0.0000  TRUE
## X-squared43   GPT-5-High vs o3-GPT-Image-High     0.090    10.3293789    0.0013  TRUE
## X-squared44   GPT-5-High vs o3-GPT-Image-Medium   0.073     3.3418006    0.0675  FALSE
## X-squared45   GPT-5-Medium vs GPT-5-Low            0.087     3.3044038    0.0691  FALSE
## X-squared46   GPT-5-Medium vs GPT-5-Minimal        0.166    12.5916034    0.0004  TRUE
## X-squared47   GPT-5-Medium vs o4-mini-High         0.051     1.4727975    0.2249  FALSE
## X-squared48   GPT-5-Medium vs o4-mini-Medium       0.088     4.6475936    0.0311  TRUE
## X-squared49   GPT-5-Medium vs o3-GPT-Image-High   0.031     0.6329902    0.4263  FALSE
## X-squared50   GPT-5-Medium vs o3-GPT-Image-Medium 0.014     0.0503367    0.8225  FALSE
## X-squared51   GPT-5-Low vs GPT-5-Minimal           0.079     2.7269186    0.0987  FALSE
## X-squared52   GPT-5-Low vs o4-mini-High           -0.036     0.6933172    0.4050  FALSE
## X-squared53   GPT-5-Low vs o4-mini-Medium         0.001     0.0000000    1.0000  FALSE
## X-squared54   GPT-5-Low vs o3-GPT-Image-High     -0.056     2.1845895    0.1394  FALSE
## X-squared55   GPT-5-Low vs o3-GPT-Image-Medium   -0.073     2.2565809    0.1330  FALSE
## X-squared56   GPT-5-Minimal vs o4-mini-High      -0.115     8.0506453    0.0045  TRUE
## X-squared57   GPT-5-Minimal vs o4-mini-Medium    -0.078     3.5912963    0.0581  FALSE
## X-squared58   GPT-5-Minimal vs o3-GPT-Image-High -0.135    13.5017616    0.0002  TRUE
## X-squared59   GPT-5-Minimal vs o3-GPT-Image-Medium -0.152  10.4699552    0.0012  TRUE
## X-squared60   o4-mini-High vs o4-mini-Medium       0.037     1.1969068    0.2739  FALSE
## X-squared61   o4-mini-High vs o3-GPT-Image-High   -0.020     0.4267414    0.5136  FALSE
## X-squared62   o4-mini-High vs o3-GPT-Image-Medium -0.037     0.7217947    0.3956  FALSE
## X-squared63   o4-mini-Medium vs o3-GPT-Image-High -0.057     3.9692535    0.0463  TRUE
## X-squared64   o4-mini-Medium vs o3-GPT-Image-Medium -0.074   3.2155324    0.0729  FALSE
```

```
## X-squared65   o3-GPT-Image-High vs o3-GPT-Image-Medium    -0.017     0.1573361    0.6916  FALSE
```

Visualization of 48 Novel Reasoning Variations

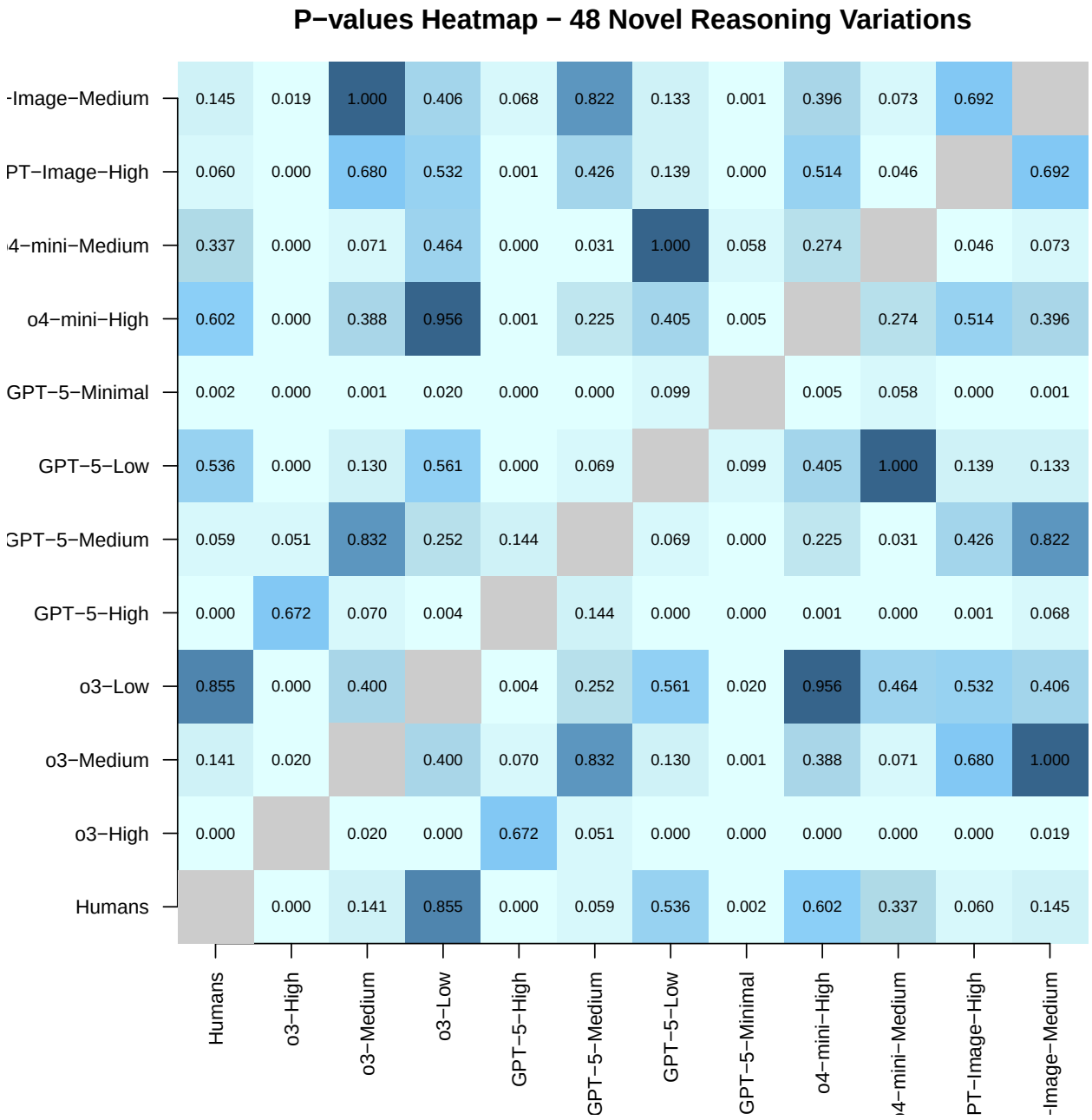
```
# Plot proportions with confidence intervals for 48 Novel reasoning variations
novel_48_reasoning_plot <- ggplot(novel_reasoning_data, aes(x = reorder(model, proportion), y = proporti
  geom_point(size = 4, color = "orange") +
  geom_errorbar(aes(ymin = proportion - 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    ymax = proportion + 1.96 * sqrt(proportion * (1 - proportion) / max_score)),
                width = 0.2, size = 1, color = "orange") +
  coord_flip() +
  theme_minimal() +
  labs(title = "48 Novel Reasoning Variations - Proportions with 95% CI",
       x = "Model",
       y = "Proportion Correct") +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))
print(novel_48_reasoning_plot)
```



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)
    }
  }
}
```



Summary of Significant Differences - 48 Novel Reasoning Variations

```
# 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:
```

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

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

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

```
##   Total comparisons: 66
```

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

```
##   Significant differences: 25
```

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

```
##   Percentage significant: 37.9 %
```

```
# 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                      -0.1370    0.0000
## X-squared3     Humans vs GPT-5-High                   -0.1234    0.0000
## X-squared6     Humans vs GPT-5-Minimal                 0.1018    0.0025
## X-squared11    o3-High vs o3-Medium                    0.0864    0.0198
## X-squared12    o3-High vs o3-Low                       0.1288    0.0005
## X-squared15    o3-High vs GPT-5-Low                    0.1595    0.0000
## X-squared16    o3-High vs GPT-5-Minimal                0.2388    0.0000
## X-squared17    o3-High vs o4-mini-High                 0.1235    0.0000
## X-squared18    o3-High vs o4-mini-Medium               0.1609    0.0000
## X-squared19    o3-High vs o3-GPT-Image-High            0.1038    0.0000
## X-squared20    o3-High vs o3-GPT-Image-Medium          0.0869    0.0190
## X-squared25    o3-Medium vs GPT-5-Minimal              0.1524    0.0012
## X-squared30    o3-Low vs GPT-5-High                   -0.1152    0.0037
## X-squared33    o3-Low vs GPT-5-Minimal                 0.1100    0.0203
## X-squared39    GPT-5-High vs GPT-5-Low                 0.1459    0.0002
## X-squared40    GPT-5-High vs GPT-5-Minimal             0.2252    0.0000
## X-squared41    GPT-5-High vs o4-mini-High              0.1099    0.0007
## X-squared42    GPT-5-High vs o4-mini-Medium            0.1473    0.0000
## X-squared43    GPT-5-High vs o3-GPT-Image-High         0.0902    0.0013
## X-squared46    GPT-5-Medium vs GPT-5-Minimal           0.1661    0.0004
## X-squared48    GPT-5-Medium vs o4-mini-Medium          0.0882    0.0311
## X-squared56    GPT-5-Minimal vs o4-mini-High          -0.1153    0.0045
## X-squared58    GPT-5-Minimal vs o3-GPT-Image-High     -0.1350    0.0002
## X-squared59    GPT-5-Minimal vs o3-GPT-Image-Medium   -0.1518    0.0012
## X-squared63    o4-mini-Medium vs o3-GPT-Image-High    -0.0571    0.0463
```

Combined Summary of Reasoning Variations

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

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

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

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

```
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:  0.541  vs  0.647
## Difference:  -0.106
## Chi-squared:  35.828
## Degrees of freedom:  1
## P-value:  0.000000002156
## 95% CI: [ -0.139 ,  -0.072 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o3-Medium
## ----------------------------------------
## Proportions:  0.541  vs  0.572
## Difference:  -0.031
## Chi-squared:  0.981
## Degrees of freedom:  1
## P-value:  0.322
## 95% CI: [ -0.09 ,  0.028 ]
## Significant:  NO
##
##  Humans vs o3-Low
## ----------------------------------------
## Proportions:  0.541  vs  0.548
## Difference:  -0.007
## Chi-squared:  0.029
## Degrees of freedom:  1
## P-value:  0.8647
## 95% CI: [ -0.066 ,  0.052 ]
## Significant:  NO
##
##  Humans vs GPT-5-High
## ----------------------------------------
## Proportions:  0.541  vs  0.668
```

```
## Difference:  -0.127
## Chi-squared:  35.763
## Degrees of freedom:  1
## P-value:  0.000000002229
## 95% CI: [ -0.167 ,  -0.087 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs GPT-5-Medium
## ----------------------------------------
## Proportions:  0.541  vs  0.594
## Difference:  -0.053
## Chi-squared:  3.045
## Degrees of freedom:  1
## P-value:  0.08098
## 95% CI: [ -0.111 ,  0.006 ]
## Significant:  NO
##
##  Humans vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.541  vs  0.51
## Difference:  0.031
## Chi-squared:  1.011
## Degrees of freedom:  1
## P-value:  0.3146
## 95% CI: [ -0.028 ,  0.091 ]
## Significant:  NO
##
##  Humans vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.541  vs  0.408
## Difference:  0.133
## Chi-squared:  19.902
## Degrees of freedom:  1
## P-value:  0.00000815
## 95% CI: [ 0.074 ,  0.191 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o4-mini-High
## ----------------------------------------
## Proportions:  0.541  vs  0.533
## Difference:  0.008
## Chi-squared:  0.106
## Degrees of freedom:  1
## P-value:  0.7448
## 95% CI: [ -0.035 ,  0.05 ]
## Significant:  NO
##
##  Humans vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.541  vs  0.489
## Difference:  0.052
## Chi-squared:  5.889
## Degrees of freedom:  1
## P-value:  0.01523
```

```
## 95% CI: [ 0.01 ,  0.095 ]
## Significant:  YES (p < 0.05)
##
##  Humans vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.541  vs  0.556
## Difference:  -0.015
## Chi-squared:  0.861
## Degrees of freedom:  1
## P-value:  0.3533
## 95% CI: [ -0.046 ,  0.016 ]
## Significant:  NO
##
##  Humans vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.541  vs  0.555
## Difference:  -0.014
## Chi-squared:  0.182
## Degrees of freedom:  1
## P-value:  0.6696
## 95% CI: [ -0.073 ,  0.045 ]
## Significant:  NO
##
##  o3-High vs o3-Medium
## ---------------------------------------
## Proportions:  0.647  vs  0.572
## Difference:  0.075
## Chi-squared:  5.078
## Degrees of freedom:  1
## P-value:  0.02423
## 95% CI: [ 0.009 ,  0.141 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o3-Low
## ---------------------------------------
## Proportions:  0.647  vs  0.548
## Difference:  0.099
## Chi-squared:  8.942
## Degrees of freedom:  1
## P-value:  0.002787
## 95% CI: [ 0.032 ,  0.166 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs GPT-5-High
## ---------------------------------------
## Proportions:  0.647  vs  0.668
## Difference:  -0.021
## Chi-squared:  0.644
## Degrees of freedom:  1
## P-value:  0.4223
## 95% CI: [ -0.072 ,  0.029 ]
## Significant:  NO
##
##  o3-High vs GPT-5-Medium
```

```
## ----------------------------------------
## Proportions:  0.647  vs  0.594
## Difference:  0.053
## Chi-squared:  2.47
## Degrees of freedom:  1
## P-value:  0.116
## 95% CI: [ -0.013 ,  0.119 ]
## Significant:  NO
##
##  o3-High vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.647  vs  0.51
## Difference:  0.137
## Chi-squared:  17.184
## Degrees of freedom:  1
## P-value:  0.00003392
## 95% CI: [ 0.07 ,  0.204 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.647  vs  0.408
## Difference:  0.238
## Chi-squared:  51.751
## Degrees of freedom:  1
## P-value:  0.00000000000063
## 95% CI: [ 0.172 ,  0.304 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o4-mini-High
## ----------------------------------------
## Proportions:  0.647  vs  0.533
## Difference:  0.113
## Chi-squared:  18.842
## Degrees of freedom:  1
## P-value:  0.0000142
## 95% CI: [ 0.061 ,  0.165 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.647  vs  0.489
## Difference:  0.158
## Chi-squared:  36.269
## Degrees of freedom:  1
## P-value:  0.000000001719
## 95% CI: [ 0.106 ,  0.21 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.647  vs  0.556
## Difference:  0.091
## Chi-squared:  17.204
```

```
## Degrees of freedom:  1
## P-value:  0.00003357
## 95% CI: [ 0.048 ,  0.134 ]
## Significant:  YES (p < 0.05)
##
##  o3-High vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.647  vs  0.555
## Difference:  0.091
## Chi-squared:  7.62
## Degrees of freedom:  1
## P-value:  0.005773
## 95% CI: [ 0.025 ,  0.158 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs o3-Low
## ---------------------------------------
## Proportions:  0.572  vs  0.548
## Difference:  0.024
## Chi-squared:  0.261
## Degrees of freedom:  1
## P-value:  0.6092
## 95% CI: [ -0.059 ,  0.107 ]
## Significant:  NO
##
##  o3-Medium vs GPT-5-High
## ---------------------------------------
## Proportions:  0.572  vs  0.668
## Difference:  -0.096
## Chi-squared:  7.606
## Degrees of freedom:  1
## P-value:  0.005817
## 95% CI: [ -0.166 ,  -0.026 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs GPT-5-Medium
## ---------------------------------------
## Proportions:  0.572  vs  0.594
## Difference:  -0.022
## Chi-squared:  0.218
## Degrees of freedom:  1
## P-value:  0.6406
## 95% CI: [ -0.104 ,  0.06 ]
## Significant:  NO
##
##  o3-Medium vs GPT-5-Low
## ---------------------------------------
## Proportions:  0.572  vs  0.51
## Difference:  0.062
## Chi-squared:  2.081
## Degrees of freedom:  1
## P-value:  0.1492
## 95% CI: [ -0.021 ,  0.145 ]
## Significant:  NO
```

```
##
##  o3-Medium vs GPT-5-Minimal
## ---------------------------------------
## Proportions:  0.572  vs  0.408
## Difference:  0.164
## Chi-squared:  15.402
## Degrees of freedom:  1
## P-value:  0.00008689
## 95% CI: [ 0.081 ,  0.246 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs o4-mini-High
## ---------------------------------------
## Proportions:  0.572  vs  0.533
## Difference:  0.039
## Chi-squared:  1.05
## Degrees of freedom:  1
## P-value:  0.3056
## 95% CI: [ -0.033 ,  0.11 ]
## Significant:  NO
##
##  o3-Medium vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.572  vs  0.489
## Difference:  0.083
## Chi-squared:  5.193
## Degrees of freedom:  1
## P-value:  0.02268
## 95% CI: [ 0.012 ,  0.154 ]
## Significant:  YES (p < 0.05)
##
##  o3-Medium vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.572  vs  0.556
## Difference:  0.016
## Chi-squared:  0.187
## Degrees of freedom:  1
## P-value:  0.6658
## 95% CI: [ -0.049 ,  0.081 ]
## Significant:  NO
##
##  o3-Medium vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.572  vs  0.555
## Difference:  0.017
## Chi-squared:  0.106
## Degrees of freedom:  1
## P-value:  0.7445
## 95% CI: [ -0.066 ,  0.099 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-High
## ---------------------------------------
## Proportions:  0.548  vs  0.668
```

```
## Difference:  -0.12
## Chi-squared:  11.896
## Degrees of freedom:  1
## P-value:  0.0005624
## 95% CI: [ -0.191 ,  -0.05 ]
## Significant:  YES (p < 0.05)
##
##  o3-Low vs GPT-5-Medium
## ---------------------------------------
## Proportions:  0.548  vs  0.594
## Difference:  -0.046
## Chi-squared:  1.124
## Degrees of freedom:  1
## P-value:  0.289
## 95% CI: [ -0.129 ,  0.036 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-Low
## ---------------------------------------
## Proportions:  0.548  vs  0.51
## Difference:  0.038
## Chi-squared:  0.722
## Degrees of freedom:  1
## P-value:  0.3953
## 95% CI: [ -0.045 ,  0.121 ]
## Significant:  NO
##
##  o3-Low vs GPT-5-Minimal
## ---------------------------------------
## Proportions:  0.548  vs  0.408
## Difference:  0.139
## Chi-squared:  11.141
## Degrees of freedom:  1
## P-value:  0.0008442
## 95% CI: [ 0.057 ,  0.222 ]
## Significant:  YES (p < 0.05)
##
##  o3-Low vs o4-mini-High
## ---------------------------------------
## Proportions:  0.548  vs  0.533
## Difference:  0.015
## Chi-squared:  0.116
## Degrees of freedom:  1
## P-value:  0.7332
## 95% CI: [ -0.057 ,  0.086 ]
## Significant:  NO
##
##  o3-Low vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.548  vs  0.489
## Difference:  0.059
## Chi-squared:  2.552
## Degrees of freedom:  1
## P-value:  0.1101
```

```
## 95% CI: [ -0.013 ,  0.131 ]
## Significant:  NO
##
##  o3-Low vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.548  vs  0.556
## Difference:  -0.008
## Chi-squared:  0.035
## Degrees of freedom:  1
## P-value:  0.8506
## 95% CI: [ -0.073 ,  0.057 ]
## Significant:  NO
##
##  o3-Low vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.548  vs  0.555
## Difference:  -0.008
## Chi-squared:  0.011
## Degrees of freedom:  1
## P-value:  0.9179
## 95% CI: [ -0.09 ,  0.075 ]
## Significant:  NO
##
##  GPT-5-High vs GPT-5-Medium
## ---------------------------------------
## Proportions:  0.668  vs  0.594
## Difference:  0.074
## Chi-squared:  4.481
## Degrees of freedom:  1
## P-value:  0.03427
## 95% CI: [ 0.005 ,  0.144 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs GPT-5-Low
## ---------------------------------------
## Proportions:  0.668  vs  0.51
## Difference:  0.158
## Chi-squared:  20.525
## Degrees of freedom:  1
## P-value:  0.000005885
## 95% CI: [ 0.088 ,  0.229 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs GPT-5-Minimal
## ---------------------------------------
## Proportions:  0.668  vs  0.408
## Difference:  0.26
## Chi-squared:  54.429
## Degrees of freedom:  1
## P-value:  0.0000000000001612
## 95% CI: [ 0.19 ,  0.33 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o4-mini-High
```

```
## ----------------------------------------
## Proportions:  0.668  vs  0.533
## Difference:  0.135
## Chi-squared:  22.201
## Degrees of freedom:  1
## P-value:  0.000002456
## 95% CI: [ 0.078 ,  0.191 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.668  vs  0.489
## Difference:  0.179
## Chi-squared:  38.844
## Degrees of freedom:  1
## P-value:  0.000000000459
## 95% CI: [ 0.123 ,  0.236 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.668  vs  0.556
## Difference:  0.112
## Chi-squared:  20.427
## Degrees of freedom:  1
## P-value:  0.000006194
## 95% CI: [ 0.064 ,  0.161 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-High vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.668  vs  0.555
## Difference:  0.113
## Chi-squared:  10.456
## Degrees of freedom:  1
## P-value:  0.001222
## 95% CI: [ 0.043 ,  0.183 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs GPT-5-Low
## ----------------------------------------
## Proportions:  0.594  vs  0.51
## Difference:  0.084
## Chi-squared:  3.962
## Degrees of freedom:  1
## P-value:  0.04654
## 95% CI: [ 0.002 ,  0.167 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.594  vs  0.408
## Difference:  0.186
## Chi-squared:  19.944
```

```
## Degrees of freedom:  1
## P-value:  0.000007975
## 95% CI: [ 0.104 ,  0.268 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs o4-mini-High
## ----------------------------------------
## Proportions:  0.594  vs  0.533
## Difference:  0.061
## Chi-squared:  2.741
## Degrees of freedom:  1
## P-value:  0.09781
## 95% CI: [ -0.01 ,  0.132 ]
## Significant:  NO
##
##  GPT-5-Medium vs o4-mini-Medium
## ----------------------------------------
## Proportions:  0.594  vs  0.489
## Difference:  0.105
## Chi-squared:  8.45
## Degrees of freedom:  1
## P-value:  0.003651
## 95% CI: [ 0.034 ,  0.176 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Medium vs o3-GPT-Image-High
## ----------------------------------------
## Proportions:  0.594  vs  0.556
## Difference:  0.038
## Chi-squared:  1.263
## Degrees of freedom:  1
## P-value:  0.2612
## 95% CI: [ -0.026 ,  0.102 ]
## Significant:  NO
##
##  GPT-5-Medium vs o3-GPT-Image-Medium
## ----------------------------------------
## Proportions:  0.594  vs  0.555
## Difference:  0.039
## Chi-squared:  0.766
## Degrees of freedom:  1
## P-value:  0.3815
## 95% CI: [ -0.044 ,  0.121 ]
## Significant:  NO
##
##  GPT-5-Low vs GPT-5-Minimal
## ----------------------------------------
## Proportions:  0.51  vs  0.408
## Difference:  0.101
## Chi-squared:  5.82
## Degrees of freedom:  1
## P-value:  0.01584
## 95% CI: [ 0.019 ,  0.184 ]
## Significant:  YES (p < 0.05)
```

```
##
##  GPT-5-Low vs o4-mini-High
## ---------------------------------------
## Proportions:  0.51  vs  0.533
## Difference:  -0.023
## Chi-squared:  0.352
## Degrees of freedom:  1
## P-value:  0.5528
## 95% CI: [ -0.095 ,  0.048 ]
## Significant:  NO
##
##  GPT-5-Low vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.51  vs  0.489
## Difference:  0.021
## Chi-squared:  0.274
## Degrees of freedom:  1
## P-value:  0.6008
## 95% CI: [ -0.051 ,  0.093 ]
## Significant:  NO
##
##  GPT-5-Low vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.51  vs  0.556
## Difference:  -0.046
## Chi-squared:  1.876
## Degrees of freedom:  1
## P-value:  0.1707
## 95% CI: [ -0.111 ,  0.019 ]
## Significant:  NO
##
##  GPT-5-Low vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.51  vs  0.555
## Difference:  -0.045
## Chi-squared:  1.071
## Degrees of freedom:  1
## P-value:  0.3007
## 95% CI: [ -0.129 ,  0.038 ]
## Significant:  NO
##
##  GPT-5-Minimal vs o4-mini-High
## ---------------------------------------
## Proportions:  0.408  vs  0.533
## Difference:  -0.125
## Chi-squared:  11.999
## Degrees of freedom:  1
## P-value:  0.0005322
## 95% CI: [ -0.196 ,  -0.054 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Minimal vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.408  vs  0.489
```

```
## Difference:  -0.08
## Chi-squared:  4.895
## Degrees of freedom:  1
## P-value:  0.02694
## 95% CI: [ -0.151 ,  -0.009 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Minimal vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.408  vs  0.556
## Difference:  -0.148
## Chi-squared:  20.383
## Degrees of freedom:  1
## P-value:  0.00000634
## 95% CI: [ -0.212 ,  -0.083 ]
## Significant:  YES (p < 0.05)
##
##  GPT-5-Minimal vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.408  vs  0.555
## Difference:  -0.147
## Chi-squared:  12.399
## Degrees of freedom:  1
## P-value:  0.0004296
## 95% CI: [ -0.229 ,  -0.065 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-High vs o4-mini-Medium
## ---------------------------------------
## Proportions:  0.533  vs  0.489
## Difference:  0.044
## Chi-squared:  2.198
## Degrees of freedom:  1
## P-value:  0.1382
## 95% CI: [ -0.014 ,  0.103 ]
## Significant:  NO
##
##  o4-mini-High vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.533  vs  0.556
## Difference:  -0.023
## Chi-squared:  0.739
## Degrees of freedom:  1
## P-value:  0.39
## 95% CI: [ -0.073 ,  0.027 ]
## Significant:  NO
##
##  o4-mini-High vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.533  vs  0.555
## Difference:  -0.022
## Chi-squared:  0.307
## Degrees of freedom:  1
## P-value:  0.5793
```

```
## 95% CI: [ -0.093 ,  0.049 ]
## Significant:  NO
##
##  o4-mini-Medium vs o3-GPT-Image-High
## ---------------------------------------
## Proportions:  0.489  vs  0.556
## Difference:  -0.067
## Chi-squared:  6.969
## Degrees of freedom:  1
## P-value:  0.008293
## 95% CI: [ -0.117 ,  -0.017 ]
## Significant:  YES (p < 0.05)
##
##  o4-mini-Medium vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.489  vs  0.555
## Difference:  -0.066
## Chi-squared:  3.278
## Degrees of freedom:  1
## P-value:  0.07023
## 95% CI: [ -0.138 ,  0.005 ]
## Significant:  NO
##
##  o3-GPT-Image-High vs o3-GPT-Image-Medium
## ---------------------------------------
## Proportions:  0.556  vs  0.555
## Difference:  0.001
## Chi-squared:  0
## Degrees of freedom:  1
## P-value:  1
## 95% CI: [ -0.063 ,  0.064 ]
## 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  significant
## ------------  ---------------------------------------  -------  ------------  --------  -----------
## X-squared     Humans vs o3-High                       -0.106    35.8276131    0.0000  TRUE
## X-squared1    Humans vs o3-Medium                     -0.031     0.9807840    0.3220  FALSE
## X-squared2    Humans vs o3-Low                        -0.007     0.0290485    0.8647  FALSE
## X-squared3    Humans vs GPT-5-High                    -0.127    35.7628507    0.0000  TRUE
## X-squared4    Humans vs GPT-5-Medium                  -0.053     3.0450645    0.0810  FALSE
```

```
## X-squared5    Humans vs GPT-5-Low                    0.031     1.0110574    0.3146  FALSE
## X-squared6    Humans vs GPT-5-Minimal                0.133    19.9024437    0.0000  TRUE
## X-squared7    Humans vs o4-mini-High                 0.008     0.1059207    0.7448  FALSE
## X-squared8    Humans vs o4-mini-Medium               0.052     5.8890900    0.0152  TRUE
## X-squared9    Humans vs o3-GPT-Image-High           -0.015     0.8614579    0.3533  FALSE
## X-squared10   Humans vs o3-GPT-Image-Medium         -0.014     0.1821217    0.6696  FALSE
## X-squared11   o3-High vs o3-Medium                   0.075     5.0782966    0.0242  TRUE
## X-squared12   o3-High vs o3-Low                      0.099     8.9417175    0.0028  TRUE
## X-squared13   o3-High vs GPT-5-High                 -0.021     0.6439982    0.4223  FALSE
## X-squared14   o3-High vs GPT-5-Medium                0.053     2.4699963    0.1160  FALSE
## X-squared15   o3-High vs GPT-5-Low                   0.137    17.1844042    0.0000  TRUE
## X-squared16   o3-High vs GPT-5-Minimal               0.238    51.7511543    0.0000  TRUE
## X-squared17   o3-High vs o4-mini-High                0.113    18.8417640    0.0000  TRUE
## X-squared18   o3-High vs o4-mini-Medium              0.158    36.2686162    0.0000  TRUE
## X-squared19   o3-High vs o3-GPT-Image-High          0.091    17.2042549    0.0000  TRUE
## X-squared20   o3-High vs o3-GPT-Image-Medium        0.091     7.6196760    0.0058  TRUE
## X-squared21   o3-Medium vs o3-Low                   0.024     0.2612695    0.6092  FALSE
## X-squared22   o3-Medium vs GPT-5-High              -0.096     7.6061759    0.0058  TRUE
## X-squared23   o3-Medium vs GPT-5-Medium            -0.022     0.2179835    0.6406  FALSE
## X-squared24   o3-Medium vs GPT-5-Low                0.062     2.0808022    0.1492  FALSE
## X-squared25   o3-Medium vs GPT-5-Minimal            0.164    15.4020623    0.0001  TRUE
## X-squared26   o3-Medium vs o4-mini-High             0.039     1.0495693    0.3056  FALSE
## X-squared27   o3-Medium vs o4-mini-Medium           0.083     5.1930769    0.0227  TRUE
## X-squared28   o3-Medium vs o3-GPT-Image-High       0.016     0.1865379    0.6658  FALSE
## X-squared29   o3-Medium vs o3-GPT-Image-Medium     0.017     0.1062288    0.7445  FALSE
## X-squared30   o3-Low vs GPT-5-High                -0.120    11.8963881    0.0006  TRUE
## X-squared31   o3-Low vs GPT-5-Medium              -0.046     1.1241354    0.2890  FALSE
## X-squared32   o3-Low vs GPT-5-Low                  0.038     0.7224506    0.3953  FALSE
## X-squared33   o3-Low vs GPT-5-Minimal              0.139    11.1413634    0.0008  TRUE
## X-squared34   o3-Low vs o4-mini-High               0.015     0.1161751    0.7332  FALSE
## X-squared35   o3-Low vs o4-mini-Medium             0.059     2.5521779    0.1101  FALSE
## X-squared36   o3-Low vs o3-GPT-Image-High         -0.008     0.0354692    0.8506  FALSE
## X-squared37   o3-Low vs o3-GPT-Image-Medium       -0.008     0.0106285    0.9179  FALSE
## X-squared38   GPT-5-High vs GPT-5-Medium            0.074     4.4813215    0.0343  TRUE
## X-squared39   GPT-5-High vs GPT-5-Low               0.158    20.5251499    0.0000  TRUE
## X-squared40   GPT-5-High vs GPT-5-Minimal           0.260    54.4285935    0.0000  TRUE
## X-squared41   GPT-5-High vs o4-mini-High            0.135    22.2007055    0.0000  TRUE
## X-squared42   GPT-5-High vs o4-mini-Medium          0.179    38.8443719    0.0000  TRUE
## X-squared43   GPT-5-High vs o3-GPT-Image-High      0.112    20.4271734    0.0000  TRUE
## X-squared44   GPT-5-High vs o3-GPT-Image-Medium    0.113    10.4560647    0.0012  TRUE
## X-squared45   GPT-5-Medium vs GPT-5-Low            0.084     3.9620963    0.0465  TRUE
## X-squared46   GPT-5-Medium vs GPT-5-Minimal        0.186    19.9438748    0.0000  TRUE
## X-squared47   GPT-5-Medium vs o4-mini-High         0.061     2.7408694    0.0978  FALSE
## X-squared48   GPT-5-Medium vs o4-mini-Medium       0.105     8.4497568    0.0037  TRUE
## X-squared49   GPT-5-Medium vs o3-GPT-Image-High   0.038     1.2626021    0.2612  FALSE
## X-squared50   GPT-5-Medium vs o3-GPT-Image-Medium 0.039     0.7659031    0.3815  FALSE
## X-squared51   GPT-5-Low vs GPT-5-Minimal           0.101     5.8203543    0.0158  TRUE
## X-squared52   GPT-5-Low vs o4-mini-High           -0.023     0.3523712    0.5528  FALSE
## X-squared53   GPT-5-Low vs o4-mini-Medium          0.021     0.2737617    0.6008  FALSE
## X-squared54   GPT-5-Low vs o3-GPT-Image-High      -0.046     1.8764212    0.1707  FALSE
## X-squared55   GPT-5-Low vs o3-GPT-Image-Medium    -0.045     1.0711198    0.3007  FALSE
## X-squared56   GPT-5-Minimal vs o4-mini-High       -0.125    11.9991492    0.0005  TRUE
## X-squared57   GPT-5-Minimal vs o4-mini-Medium     -0.080     4.8949612    0.0269  TRUE
## X-squared58   GPT-5-Minimal vs o3-GPT-Image-High  -0.148    20.3827025    0.0000  TRUE
```

```
## X-squared59   GPT-5-Minimal vs o3-GPT-Image-Medium        -0.147    12.3986657    0.0004  TRUE
## X-squared60   o4-mini-High vs o4-mini-Medium               0.044     2.1983980    0.1382  FALSE
## X-squared61   o4-mini-High vs o3-GPT-Image-High           -0.023     0.7390554    0.3900  FALSE
## X-squared62   o4-mini-High vs o3-GPT-Image-Medium         -0.022     0.3073297    0.5793  FALSE
## X-squared63   o4-mini-Medium vs o3-GPT-Image-High         -0.067     6.9691144    0.0083  TRUE
## X-squared64   o4-mini-Medium vs o3-GPT-Image-Medium       -0.066     3.2777315    0.0702  FALSE
## X-squared65   o3-GPT-Image-High vs o3-GPT-Image-Medium     0.001     0.0000000    1.0000  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: 66
```

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

```
##   Significant differences: 33
```

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

```
##   Percentage significant: 50 %
```

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

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

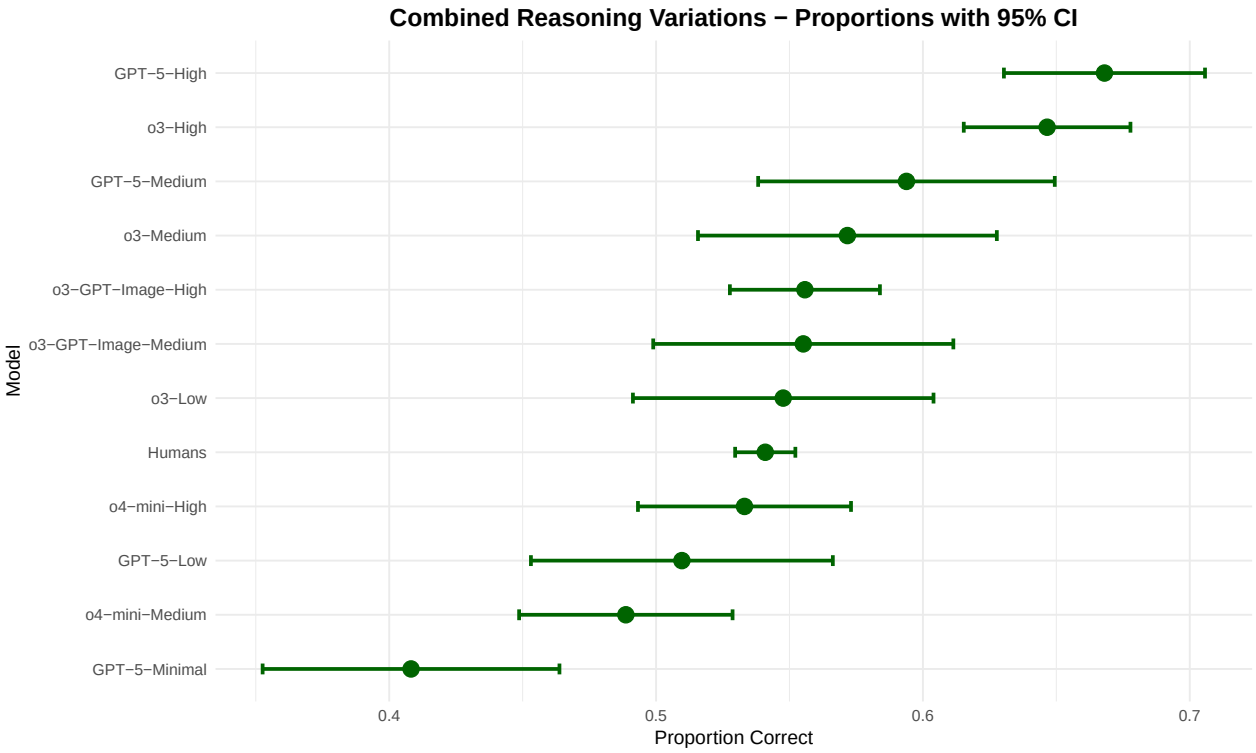
```
combined_reasoning_sig <- combined_reasoning_results[combined_reasoning_results$significant, c("compari
if (nrow(combined_reasoning_sig) > 0) {
  print(kable(combined_reasoning_sig, format = "simple", digits = 4))
} else {
  cat("  None\n")
}
```

```
##
##
##               comparison                              diff   p_value
## ------------  ------------------------------------  --------  --------
## X-squared     Humans vs o3-High                      -0.1056    0.0000
## X-squared3    Humans vs GPT-5-High                   -0.1271    0.0000
## X-squared6    Humans vs GPT-5-Minimal                 0.1327    0.0000
## X-squared8    Humans vs o4-mini-Medium                0.0522    0.0152
## X-squared11   o3-High vs o3-Medium                    0.0748    0.0242
## X-squared12   o3-High vs o3-Low                       0.0989    0.0028
## X-squared15   o3-High vs GPT-5-Low                    0.1369    0.0000
## X-squared16   o3-High vs GPT-5-Minimal                0.2384    0.0000
## X-squared17   o3-High vs o4-mini-High                 0.1134    0.0000
## X-squared18   o3-High vs o4-mini-Medium               0.1579    0.0000
## X-squared19   o3-High vs o3-GPT-Image-High            0.0908    0.0000
## X-squared20   o3-High vs o3-GPT-Image-Medium          0.0914    0.0058
## X-squared22   o3-Medium vs GPT-5-High                -0.0963    0.0058
## X-squared25   o3-Medium vs GPT-5-Minimal              0.1635    0.0001
## X-squared27   o3-Medium vs o4-mini-Medium             0.0830    0.0227
```

```
## X-squared30   o3-Low vs GPT-5-High                    -0.1204    0.0006
## X-squared33   o3-Low vs GPT-5-Minimal                  0.1395    0.0008
## X-squared38   GPT-5-High vs GPT-5-Medium               0.0742    0.0343
## X-squared39   GPT-5-High vs GPT-5-Low                  0.1584    0.0000
## X-squared40   GPT-5-High vs GPT-5-Minimal              0.2599    0.0000
## X-squared41   GPT-5-High vs o4-mini-High               0.1349    0.0000
## X-squared42   GPT-5-High vs o4-mini-Medium             0.1794    0.0000
## X-squared43   GPT-5-High vs o3-GPT-Image-High          0.1123    0.0000
## X-squared44   GPT-5-High vs o3-GPT-Image-Medium        0.1129    0.0012
## X-squared45   GPT-5-Medium vs GPT-5-Low                0.0842    0.0465
## X-squared46   GPT-5-Medium vs GPT-5-Minimal            0.1857    0.0000
## X-squared48   GPT-5-Medium vs o4-mini-Medium          0.1052    0.0037
## X-squared51   GPT-5-Low vs GPT-5-Minimal              0.1015    0.0158
## X-squared56   GPT-5-Minimal vs o4-mini-High          -0.1250    0.0005
## X-squared57   GPT-5-Minimal vs o4-mini-Medium        -0.0805    0.0269
## X-squared58   GPT-5-Minimal vs o3-GPT-Image-High     -0.1476    0.0000
## X-squared59   GPT-5-Minimal vs o3-GPT-Image-Medium   -0.1470    0.0004
## X-squared63   o4-mini-Medium vs o3-GPT-Image-High    -0.0671    0.0083
```

Visualization of Combined Reasoning Variations

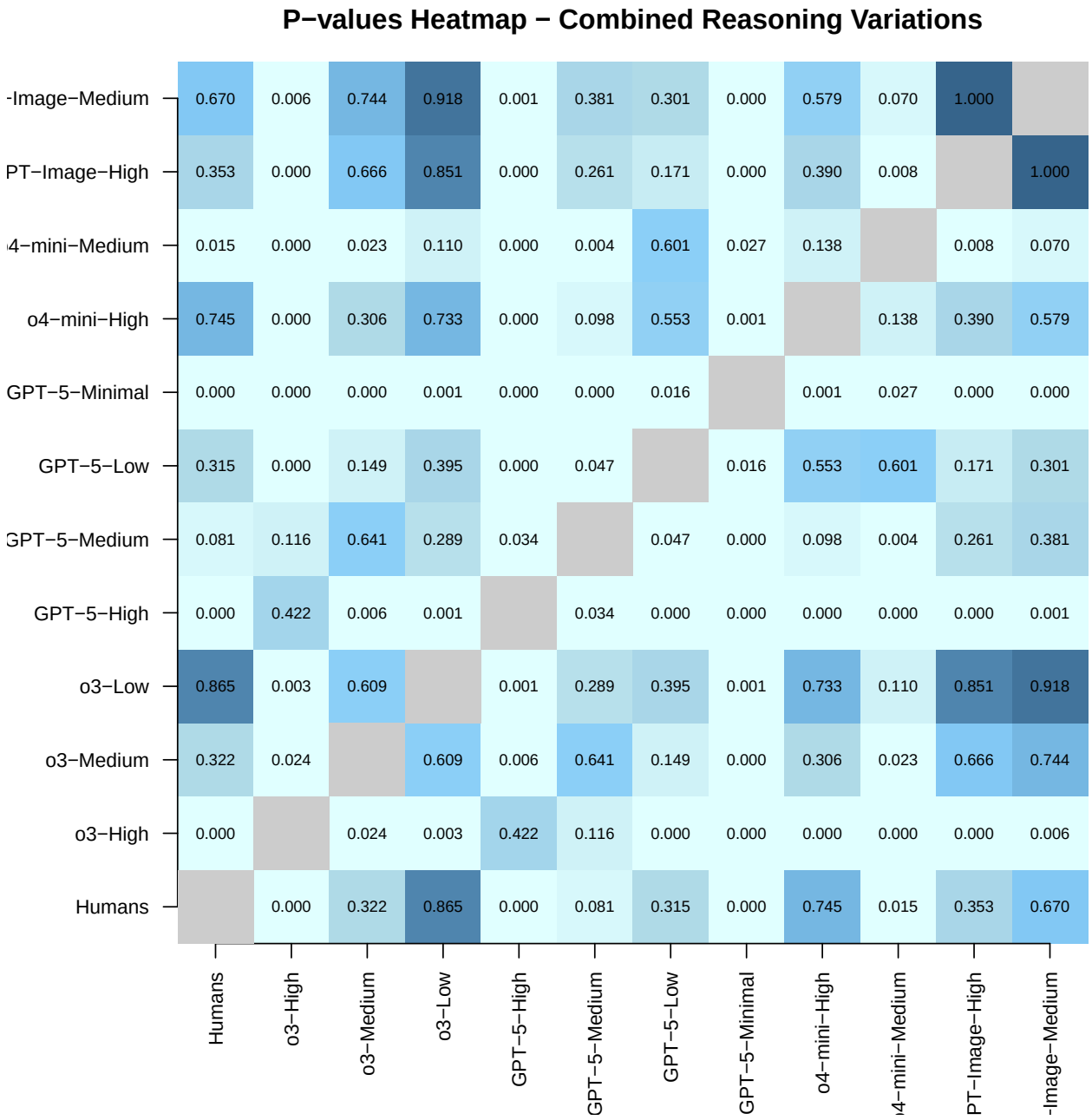
```r
# Plot proportions with confidence intervals for combined reasoning variations
combined_reasoning_plot <- ggplot(collapsed_reasoning_data, aes(x = reorder(model, proportion), y = prop
  geom_point(size = 4, color = "darkgreen") +
  geom_errorbar(aes(ymin = proportion - 1.96 * sqrt(proportion * (1 - proportion) / max_score),
                    ymax = proportion + 1.96 * sqrt(proportion * (1 - proportion) / max_score)),
                width = 0.2, size = 1, color = "darkgreen") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Combined Reasoning Variations - Proportions with 95% CI",
       x = "Model",
       y = "Proportion Correct") +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))
print(combined_reasoning_plot)
```



Heatmap for Combined Reasoning Variations

```r
# 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(co
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_mod
     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_mod
     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)
    }
  }
}
```



Summary of Significant Differences - Combined Reasoning Variations

```
# 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:
```

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

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

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

```
##   Total comparisons: 66
```

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

```
##   Significant differences: 33
```

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

```
##   Percentage significant: 50 %
```

```
# Show which comparisons are significant
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                      -0.1056    0.0000
## X-squared3     Humans vs GPT-5-High                   -0.1271    0.0000
## X-squared6     Humans vs GPT-5-Minimal                 0.1327    0.0000
## X-squared8     Humans vs o4-mini-Medium                0.0522    0.0152
## X-squared11    o3-High vs o3-Medium                    0.0748    0.0242
## X-squared12    o3-High vs o3-Low                       0.0989    0.0028
## X-squared15    o3-High vs GPT-5-Low                    0.1369    0.0000
## X-squared16    o3-High vs GPT-5-Minimal                0.2384    0.0000
## X-squared17    o3-High vs o4-mini-High                 0.1134    0.0000
## X-squared18    o3-High vs o4-mini-Medium               0.1579    0.0000
## X-squared19    o3-High vs o3-GPT-Image-High            0.0908    0.0000
## X-squared20    o3-High vs o3-GPT-Image-Medium          0.0914    0.0058
## X-squared22    o3-Medium vs GPT-5-High                -0.0963    0.0058
## X-squared25    o3-Medium vs GPT-5-Minimal              0.1635    0.0001
## X-squared27    o3-Medium vs o4-mini-Medium             0.0830    0.0227
## X-squared30    o3-Low vs GPT-5-High                   -0.1204    0.0006
## X-squared33    o3-Low vs GPT-5-Minimal                 0.1395    0.0008
## X-squared38    GPT-5-High vs GPT-5-Medium              0.0742    0.0343
## X-squared39    GPT-5-High vs GPT-5-Low                 0.1584    0.0000
## X-squared40    GPT-5-High vs GPT-5-Minimal             0.2599    0.0000
## X-squared41    GPT-5-High vs o4-mini-High              0.1349    0.0000
## X-squared42    GPT-5-High vs o4-mini-Medium            0.1794    0.0000
## X-squared43    GPT-5-High vs o3-GPT-Image-High         0.1123    0.0000
## X-squared44    GPT-5-High vs o3-GPT-Image-Medium       0.1129    0.0012
## X-squared45    GPT-5-Medium vs GPT-5-Low               0.0842    0.0465
## X-squared46    GPT-5-Medium vs GPT-5-Minimal           0.1857    0.0000
## X-squared48    GPT-5-Medium vs o4-mini-Medium          0.1052    0.0037
## X-squared51    GPT-5-Low vs GPT-5-Minimal              0.1015    0.0158
## X-squared56    GPT-5-Minimal vs o4-mini-High          -0.1250    0.0005
## X-squared57    GPT-5-Minimal vs o4-mini-Medium        -0.0805    0.0269
## X-squared58    GPT-5-Minimal vs o3-GPT-Image-High     -0.1476    0.0000
```

```
## X-squared59   GPT-5-Minimal vs o3-GPT-Image-Medium    -0.1470    0.0004
## X-squared63   o4-mini-Medium vs o3-GPT-Image-High     -0.0671    0.0083
```

Export Results to CSV

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

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

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

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

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

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