Omni-Math (n=800)

Left panel: scatter plot with x-axis "LLM compare" (ranging from approximately -0.8 to 0.4) and y-axis "Human labels" (ranging from 2 to 10). Legend: OpenAI o3, Gemini 2.5 Pro.

Right panel: y-axis "Correlation subsample" (0.6 to 0.8). Categories OpenAI o3 and Gemini 2.5 Pro. Legend: Pearson, Spearman, Kendall.