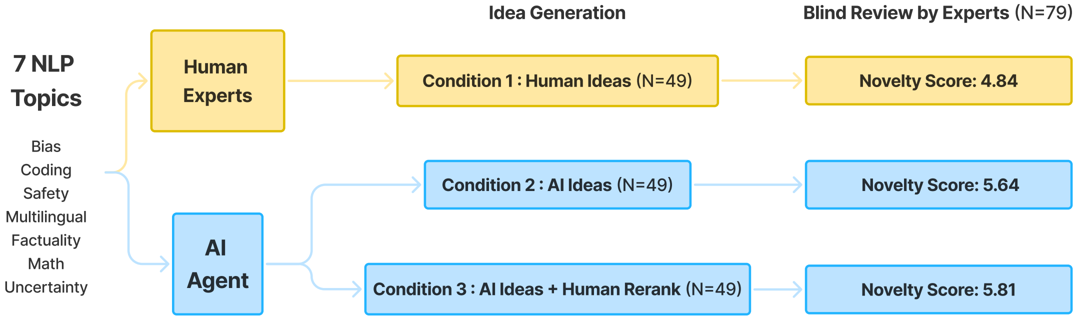

Figure 1: Overview of our study: we recruit 79 expert researchers to perform blind review of 49 ideas from each of the three conditions: expert-written ideas, AI-generated ideas, and AI-generated ideas reranked by a human expert. We standardize the format and style of ideas from all conditions before the blind review. We find AI ideas are judged as significantly more novel than human ideas ($p < 0.05$).

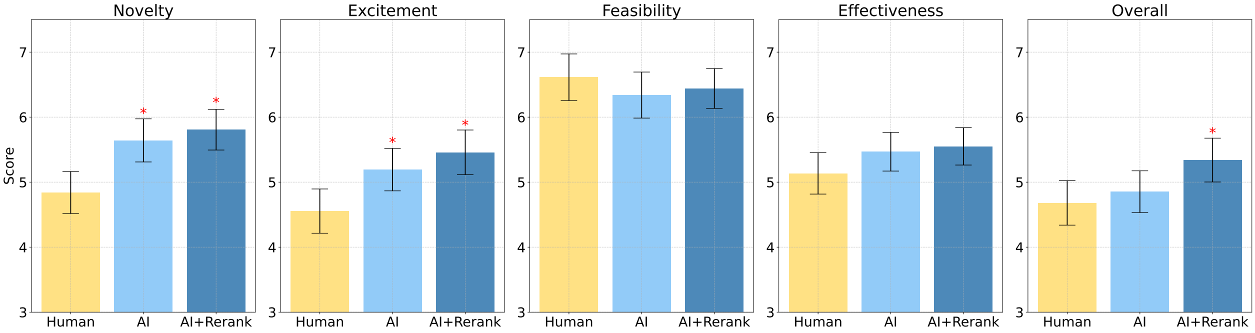

Figure 2: Comparison of the three experiment conditions across all review metrics. Red asterisks indicate that the condition is statistically better than the `Human` baseline with two-tailed Welch's t-tests and Bonferroni correction. All scores are on a 1 to 10 scale. More detailed results are in Section 5.