Appendix

A Experimental task

We report below the problem statements presented to participants during the hackathon. These statements reflected real business challenges that the respective business units were actively working on at the time of the experiment. Each statement was accompanied by relevant market data and additional contextual information provided by the business units. All statements represented significant innovation opportunities identified by senior management. For confidentiality, we have removed specific brand names and company references, indicated by [brand] or [company].

1. **Business Unit 1 Problem Statement:**
   "How to help consumers transition from product form X to Y [specific product examples removed]?"

2. **Business Unit 2 Problem Statement:**
   "How to motivate consumers who have never tried product form X to try it as part of their regimen"

3. **Business Unit 3 Problem Statement:**
   "How do we make the current portfolio of Brand X form/regimen offerings in the category simple to understand and choose to shop as a 'one size fits all solution', versus competitors who offer only a single offering?[company and competitor examples removed])?"

4. **Business Unit 4 Problem Statement:**
   "What are ways we can affect the consumer dosing habits of product X to help them achieve better health?"

Participants were assigned to work on the above-mentioned problems from their own business unit, ensuring that the task reflected their domain knowledge. Each participant (or team) followed P&G's standardized new product development process. At the end of the session, they submitted a final deliverable: a 1–2 page written solution describing their proposed new product concept, which served as the unit of analysis for our subsequent expert evaluations.

B Solution Evaluation Process

This section details the evaluation process used to assess the quality and characteristics of solutions generated during the experiment.

B.1 Evaluator Selection and Composition

The evaluation of solutions was conducted by a panel of 19 expert evaluators who collectively performed 1,595 evaluations across 550 unique solutions, resulting in an average of around three evaluations per solution. All evaluators were experienced professionals with backgrounds in business and technology - MBA and Engineering students, or recent graduates, at a top business or engineering school, ensuring a comprehensive assessment of both technical and commercial aspects of the proposed solutions.

B.2 Evaluation Process

Each evaluator was assigned approximately 70 solutions to review. For each solution, evaluators assessed the solutions, comprising five key components:

1. Idea Name
2. Recommended Solution
3. Rationale Details
4. Critical Work Required
5. Support or Resources Needed for Implementation

B.3 Evaluation Metrics

Evaluators assessed each solution on five primary dimensions using a 1-10 scale:

- **Overall Quality**: A comprehensive assessment of the solution's merit
- **Novelty**: The originality and uniqueness of the approach
- **Impact**: The effectiveness in addressing the problem and creating value
- **Business Potential**: The potential for significant business benefit and value creation
- **Feasibility**: The practicality and achievability of the proposed approach

Additionally, evaluators assessed the technical versus commercial orientation of each solution on a separate 1-7 Likert scale.

B.4 Quality Control and Interrater Reliability

To maintain evaluation integrity, evaluators agreed to strict confidentiality requirements.

To ensure evaluation reliability, solutions received multiple independent assessments. Final scores for each solution were calculated by averaging all individual evaluations.

To assess evaluation consistency, we measured inter-rater reliability using multiple metrics. Our analysis revealed an ICC2 of 0.452. We additionally report Gwet's AC2, which we prefer over AC1 because it accommodates weighted agreement on ordinal scales. To reflect that large disagreements are more

Figure A1: **Gwet's AC2**

Metric	Value
Coefficient Value	0.66
Coefficient Name	AC2
95% Confidence Interval	(0.62, 0.69)
p_value	0
Z-Score	34.04
Standard Error (SE)	0.02
Observed Agreement (Pa)	0.92
Expected Agreement (Pe)	0.76

consequential than small ones, we apply quadratic weights. As shown in Figure A1, the coefficient is 0.66, indicating high agreement.

Our approach of using 19 domain experts who conducted 1,595 evaluations across 550 solutions (averaging 2.89 assessments per solution) follows standard practice in innovation evaluation. Evaluators were blind to experimental conditions and used predefined metrics. While perfect agreement is rare in subjective, knowledge-intensive tasks, our reliability metrics provided sufficient consensus for meaningful comparison across conditions, consistent with research showing that even with moderate agreement levels, averaged ratings effectively identify quality differences (Cole et al., 1981; Wessely, 1998).

Our evaluators were MBA and engineering students who received standardized training on P&G's innovation criteria. While this approach provides consistency and external perspective appropriate for early-stage screening, it may not capture nuanced organizational or market considerations that internal experts would recognize.

We report the overall distribution of scores in Figure A2. The histogram shows substantial variation in evaluator ratings, with a roughly unimodal shape and no evidence of ceiling or floor effects. This distribution supports the use of continuous quality scores in our analyses. We report the distribution by grader in Figure A3. The KDE plots illustrate that graders vary somewhat in their average scoring levels, but their distributions substantially overlap.

Finally, we report below a table of summary statistics by grader. These statistics highlight the extent to which raters differed in their average scoring levels and use of the scale, helping explain why agreement measures that penalize systematic level differences (such as ICC) produce lower values in our setting. At the same time, the consistency of within-rater variability supports the use of agreement metrics, like Gwet's AC2, that are robust to these systematic shifts.

Figure A2: **Rating Histogram**

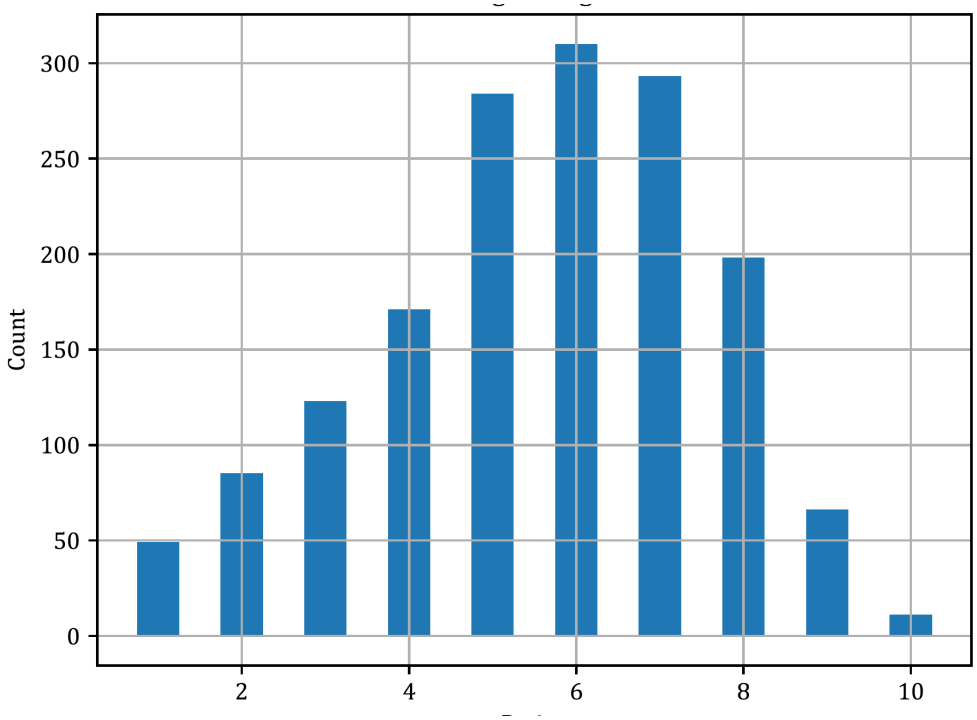


Figure A3: **KDE of Ratings by Rater**

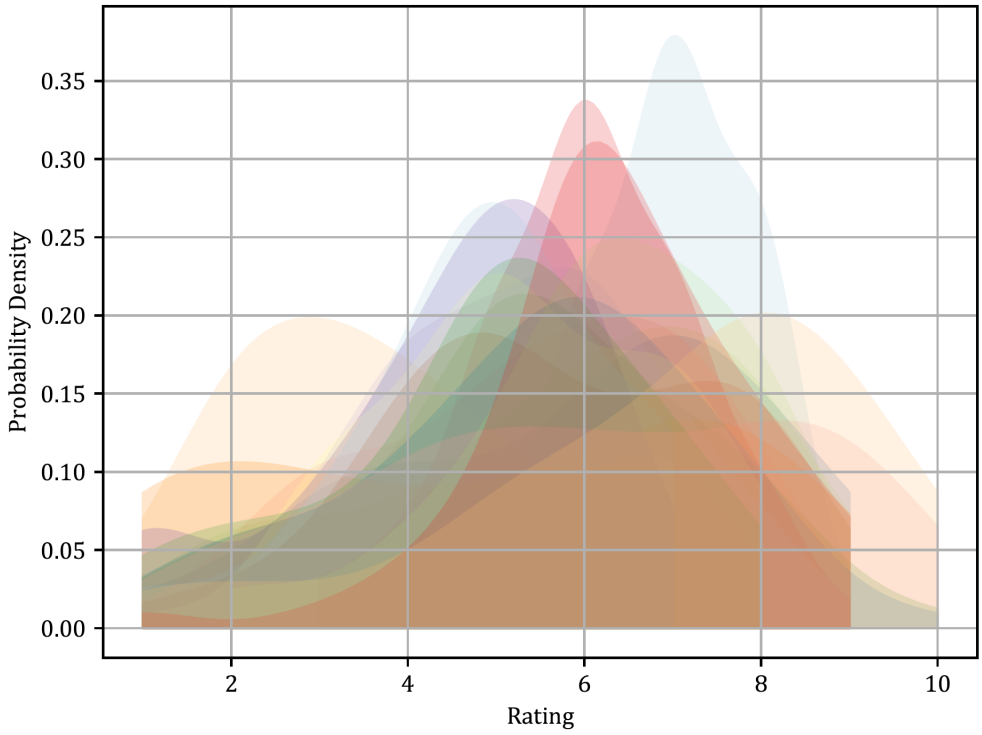


Table A1: Summary Statistics by Grader

Grader	Count	Mean	Std. Dev.	Min	Max
1	21	4.95	1.5	1	7
2	71	4.97	1.8	1	8
3	96	5.33	1.59	1	8
4	95	5.82	2.07	1	9
5	95	6.02	1.21	3	9
6	122	5.43	1.92	1	10
7	73	6.08	2.38	1	10
8	92	6.3	1.57	2	9
9	95	5.4	1.93	1	10
10	69	6.84	1.08	4	9
11	66	5.71	1.85	1	9
12	114	4.06	1.88	1	8
13	102	5.48	1.81	1	9
14	92	4.97	2.65	1	9
15	60	5.93	2.11	1	9
16	74	4.51	1.59	1	7
17	91	6.98	2.21	1	10
18	91	5.38	1.5	2	9
19	71	6.41	1.43	1	9

C Robustness Check: Controlling for Typographical Errors

To rule out the possibility that the observed performance gains are driven by cleaner writing or stylistic polish rather than by stronger ideas, we quantified the extent of surface-level writing issues in each final solution. We used the Python package `language_tool_python`, which provides access to LanguageTool's grammar and spell-checking engine. LanguageTool is a widely used open-source system capable of identifying a broad set of textual problems—such as missing punctuation, capitalization errors, grammatical inconsistencies, and redundant wording. For every solution, we computed a measure capturing the total number of issues detected by the tool, which we include as a control in our regression models.

Table A2 presents our main quality regressions with this control included. As expected, typographical errors are negatively associated with quality ratings (coefficient = -0.004, $p < 0.05$ in Model 1; coefficient = -0.004, $p < 0.05$ in Model 2). However, controlling for typos does not meaningfully change our treatment effects. All coefficients remain statistically significant and substantively similar in magnitude across all three model specifications.

In Model 1 (with typos control), Team No AI shows a 0.237 standard deviation improvement ($p < 0.05$), Individual + AI shows a 0.360 standard deviation improvement ($p < 0.01$), and Team + AI shows a 0.381 standard deviation improvement ($p < 0.01$) over the baseline. These estimates remain stable in Model 2 (with fixed effects and typos) and Model 3 (with both fixed effects and full controls, including the typo count).

The persistence of treatment effects after controlling for presentation quality provides additional evidence that our results reflect genuine differences in idea quality rather than merely better presentation or fewer errors. This finding complements our decomposition analysis in Section 5.4, which shows that AI substantially enhances the average quality of the five initial ideas generated (before participants select and develop one into a final solution), further indicating that AI's benefits stem from generating higher-quality concepts from the outset.

D Prompts

For this paper, the authors focused on creating specific prompts to integrate with the innovation process, rather than replacing it with automated systems. Our intent was not to automate any part of the existing workflow but rather to help participants engage in their standard exploratory process, using the AI as they saw fit. Rather than optimizing for precision or consistent outputs, we designed the prompts to encourage dialogue and draw out participants' assessment of the AI's outputs.

We identified specific integration points in this early innovation workflow that were both challenging and time-consuming for humans and yet straightforward for the AI, and we aimed to maximize each party's strengths. Our prompting approach integrated three elements: established business methodologies, evidence-based prompting techniques, and deliberate strategies to draw out iterative engagement and domain expertise. Prompting techniques included direct, explicit instructions, personas, clear constraints, few-shot examples, and Chain-of-Thought reasoning. Below we describe these approaches:

D.1 Chain-of-Thought

Chain-of-Thought is an established prompting technique that instructs the AI to articulate its reasoning step by step before delivering a response. This approach often involves breaking down complex tasks into smaller sequential components and asking the AI to refine its responses. We explicitly structured our prompts to mirror expert thought processes, breaking down complex tasks for better performance. For instance, in our ideation prompts, we first asked the AI to output numerous ideas and then asked it to refine and narrow down those ideas, explaining its reasoning at each step.

D.2 Purposeful Elicitation

Purposeful Elicitation involves directing the AI to ask the user questions. This technique has significant user experience implications and, in our prompts, serves three purposes. First, it makes for a longer conversation, which can improve output. In some cases, we direct the AI to ask the participant open-ended questions so that what might have been a short interaction turns into a longer conversation allowing the participant to guide output, provide more context, or redirect the conversation. Second, it helps the AI gather context and third, it can create deliberate opportunities for participant input. Creating deliberate pause points in which the AI cannot proceed without gathering information from the participants gives participants an opportunity to add their judgment or expertise to the conversation.

D.3 Personas

Personas involve assigning the AI a professional role ("you are an innovation specialist") to provide context and shape how it analyzes problems and structures responses.

D.4 Role-Play

Role-Play extends beyond persona to create interactive and dialogue-based simulations. The AI actively embodies a character (such as a simulated customer) and responds to questions, adapting its response based on the interaction. It can do so fairly realistically, even with just a prompt. The AI's ability to role-play creates a low-stakes environment for testing ideas, exploring perspectives, and following up on interesting responses that would be costly and hard to scale with real users.

D.5 Constraints

Constraints in prompts can serve as guardrails that keep the AI on track. These are not merely limitations but directives that help the AI achieve its goal. We add constraints to prompts to ensure consistency, draw out participant expertise, and to allow for natural dialogue. For instance, we instruct the AI not to "provide a solution" in the framing prompt so that participants can spend time analyzing options; we instruct the AI to only ask "one question at a time" to allow a more natural flow to the conversation, and we instruct the AI to "Wait for the team to respond. Do not move on until the team responds" in the role-play prompt so that participants and not the AI pick a specific persona to interview. Collectively, constraints can create more productive interactions, elicit participant expertise, and prevent the AI from defaulting to providing immediate solutions.

Specific prompts use these approaches in different ways.

D.6 Specific Prompts

D.6.1 Ideation Prompts

We developed ideation prompts based on well-known ideation principles including generating many ideas before evaluation, using constraints to focus the problem space, and the integration of different perspectives. The prompts begin with explicit instructions for participants to share their problem statement, followed by a structured ideation using step-by-step prompting. The prompt instructs the AI to generate many ideas and then evaluate these and modify and finally to develop each into a detailed concept. Participants can see the ideas being developed, observe evaluations, and intervene or redirect at any point.

D.6.2 Framing Prompt

Our framing prompts were built on problem-framing techniques that allow practitioners to view challenges from multiple perspectives. The Alternative Structuring of the Problem prompt establishes a persona (an innovation specialist) who guides participants through the process but whose role is constrained (analyze, but do not provide a solution). We used a few-shot approach providing examples of different frameworks without constraining the possible perspectives. The prompt was explicitly structured to create a collaborative analysis process, beginning with an introduction, an explanation of the value of reframing, and an offer to help participants view the problem from multiple perspectives.

D.6.3 Simulated Customer Interview Prompt

For customer interviews, we combined traditional market research in the form of the customer interview with the AI's capacity to role-play different personas simultaneously and quickly create numerous opportunities for simulated customer interviews. This structured prompt moves through distinct phases: persona creation, question development, interview, and post-interview analysis. The prompt establishes the AI as both a consumer psychologist (facilitator or guide) and a customer (interviewee) with clear rules about role adherence. We also create deliberate pause points requiring participant input and turn-taking and instruct the AI to encourage iteration ("do this several times with different customers") and reflection.

Prompts are provided below. Not all prompts can be provided because some are based on the proprietary processes used at the research site.

D.7 Prompts

D.7.1 Problem Definition

Basic Research

```
You are an incredibly smart and experienced research assistant asked to gather
    information to help analyze the following problem: [Insert Problem Statement]
First introduce yourself to the team and let them know that you want to help the team
    begin their research process.
Second ask them for any documents they might have to help you with research.
Then ask the team a series of questions 2-3 about the problem (ask them 1 at a time and
    wait for a response). You can also suggest responses or offer up multiple-choice
    responses if appropriate; if applicable, provide an all or none of the above option.
The goal is to narrow down your research focus. Then gather what information you can to
    try and answer those questions using the documents and what you know. Actually do
    it. Dont just say youll do it. You can also suggest other avenues for exploration to
    help analyze the problem.
```

Consumer Simulation

```
For five different consumers that have [Insert PROBLEM] provide the following in a
    succinct way:
 Describe your consumer (WHO) and their Job To Be Done (JTBD), Problem to Solve (WHAT)
 Describe the consumers current habit & how they solve the problem today.
```

Alternative Structuring of the Problem

```
You are an innovation specialist and helping a team work on the following problem:
    <INSERT PROBLEM> First introduce yourself to the team and let them know that you are
    here to help them analyze the problem. Explain that reframing a problem can be
    helpful because it can help shift the focus and help the team look at the problem
    from different angles and because it can encourage creative thinking. Then, given
    the framing of this problem, suggest 3 to 4 different ways to frame the problem.
    These can include 2x2 graphs, Porter's Five Forces, Root Cause Analysis, the 3 Ps
    for positive psychology, and more. Number those and actually frame the problem in
    italics within the frame. Tell the team they can pick any framing they like and work
    through this with you. You should work with the team, ask questions, make
    suggestions, and help them analyze this problem. Your role is not to find a solution
    but to analyze the problem.
```

D.7.2 Ideation

General Ideation

```
Generate new product ideas with the following requirements: [Insert problem statement].
    The ideas are just ideas. The product need not yet exist, nor may it necessarily be
    clearly feasible.
Follow these steps. Do each step, even if you think you do not need to. First, generate
    a list of 20 ideas (short title only). Second, go through the list and determine
    whether the ideas are different and bold, modify the ideas as needed to make them
    bolder and more different. No two ideas should be the same. This is important! Next,
    give the ideas a name and combine it with a product description. The name and idea
    are separated by a colon and followed by a description. The idea should be expressed
    as a paragraph of 40-80 words.
Do this step by step!
```

Five Vectors

```
Generate new product ideas for [INSERT PROBLEM] using the 5 vectors of superiority from
    P&G. The vectors are: Superior Product, Superior Packaging, Superior Brand
    Communication, Superior Retail Execution, and Superior Customer and Consumer Value.
    Generate 5 ideas for each vector. No ideas should be the same.
```

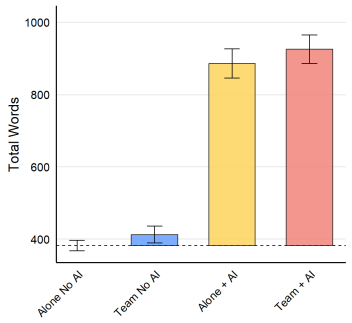
Constrained Ideation

```
Pick 4 random numbers between 1 and 11. Then, for each number, look at the appropriate
    lines on the list below and use the constraint you find for that number to generate
    an additional 3 ideas that solve the question but adhere to the constraints. Take
    the constraint literally.
List:
1 Must rhyme
2 Must be expensive
3 Must be very cheap
4 Must be very complicated
5 Must be usable by an astronaut
6 Must be usable by a superhero
7 Must be very simple
8 Must appeal to a child
9 Must be scary
10 Must be related to a book or movie
11 Must be made only of natural products
```

Selection

```
Read all the ideas so far. Select the ten ideas that combine feasibility, uniqueness,
    and the ability to drive a competitive advantage for the company the most, and
    present a chart showing the ideas and how they rank.
For each idea in the chart, describe the main features and functionalities of the
    proposed solution and how we might drive category growth (i.e., # of users, usage
    occasions, premiumization).
```

Figure A4: **Length of Solutions Produced**



Notes: This figure compares the length of solutions produced by AI-treated groups with those produced by non-AI-treated groups with standard errors.

Table A2: Solution Quality (Standardized, Controlling for Typos

	Quality	Quality	Quality
Team No AI	0.237**	0.254**	0.296**
	(0.118)	(0.120)	(0.130)
Individual + AI	0.360***	0.370***	0.355***
	(0.105)	(0.106)	(0.106)
Team + AI	0.381***	0.392***	0.446***
	(0.122)	(0.123)	(0.139)
Number of Typos	−0.004**	−0.004**	−0.003**
	(0.002)	(0.002)	(0.002)
Team+AI = Team No AI	$p = 0.238$	$p = 0.254$	$p = 0.283$
Fixed Effects		X	X
Controls			X
Control Mean	0.043	−0.110	0.344+
	(0.081)	(0.171)	(0.230)
Observations	550	550	550
Adjusted R^2	0.038	0.049	0.055

Note: Robust standard errors in parentheses. Number of Typos measured using Python's language_tool_python library. Fixed effects include date and business unit indicators. Controls include band level, years of experience, gender, and prior AI usage as discussed in the text. + $p < 0.2$, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A3: Solution Quality with Wild Cluster Bootstrap Inference

	Model 1	Model 2	Model 3
Team No AI	0.245*	0.262*	0.307**
	(0.120)	(0.122)	(0.131)
	[0.055]	[0.055]	[0.016]
Individual + AI	0.373***	0.386***	0.370***
	(0.106)	(0.108)	(0.107)
	[<0.004]	[0.004]	[<0.004]
Team + AI	0.392**	0.404**	0.463***
	(0.122)	(0.123)	(0.139)
	[0.023]	[0.031]	[0.008]
Fixed Effects		X	X
Controls			X
Observations	550	550	550
Clusters	8	8	8
R-squared	0.028	0.039	0.074

Note: Robust standard errors in parentheses. Wild cluster bootstrap p-values in brackets, computed using the Rademacher distribution with all $2^8 = 256$ possible draws. With only 8 clusters (randomization strata defined by business unit × geography), wild cluster bootstrap provides more reliable inference than conventional cluster-robust methods (Cameron et al., 2008). Fixed effects include business unit and date of participation. Controls include band level, years of experience, gender, and prior AI usage. + $p < 0.2$, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$