Asking the Right Question:

How Robot Elicitation Strategies Shape Engagement and Substantive Contribution in Creative Group Ideation

Bogdan-Andrei Micu

Supervisors: Catharine Oertel, Ruben Weijers
EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 19, 2026

Name of the student: Bogdan-Andrei Micu
Final project course: CSE3000 Research Project
Thesis committee: Catharine Oertel, Ruben Weijers, Guohao Lan

Abstract

Social robots can shape group interaction, but robot facilitation is often studied at the level of the robot as a whole rather than at the level of the specific utterances through which the robot intervenes. This paper investigates how three spoken elicitation strategies delivered by a social robot -generative, elaborative, and perspective-shifting prompts- shape participant engagement and contribution substantiveness in a two-person creative ideation task about improving the TU Delft campus experience. The study used a within-subjects design in which each of the 20 groups received one prompt from each strategy during divergence and one during convergence, while the task, robot, scheduling rule, and technical setup were kept constant. Engagement was measured using self-reports, response delay, speaking time, vocal activation, and connection cues; contribution substantiveness was measured using manually coded idea count, elaboration units, and consecutive same-subject turns. The descriptive results suggest that convergence produced higher self-reported engagement than divergence, generative prompts were most associated with idea breadth, and elaborative prompts were most associated with developed and sustained discussion. Perspective-shifting prompts appeared more useful once participants already had ideas to evaluate. The findings do not establish a universally superior strategy, but they show that robot facilitation should be designed as phase-sensitive prompt behaviour rather than treated as a general effect of robot presence.

1 Introduction

1.1 Motivation and research gap

Creative group work is a common part of education, design, and project-based collaboration, but productive ideation is difficult to support. Groups need to generate alternatives, listen to one another, develop promising directions, and eventually narrow the space of options. Prior work shows why this is fragile: participants can block one another's idea production, but exposure to others' ideas can also stimulate further search when the interaction is well supported (Nijstad and Stroebe, 2006). Facilitation therefore matters because small interventions can change whether a group produces more ideas, develops ideas more deeply, or shifts too early toward evaluation.

Social robots are promising in this setting because they can be physically present in a shared group interaction while delivering consistent interventions. Research on robots in groups shows that robots can affect coordination, participation, and the way people respond to one another (Sebo et al., 2020). Recent work frames this as interaction-shaping robotics: robots can be studied by how they influence interactions between other agents, not only by how one person interacts with the robot (Gillet et al., 2024). For creative ideation, a robot facilitator does not need to be an idea-generating teammate; it can instead ask a short question that may shift what the human participants do next.

However, prior findings also show that robot facilitation is not automatically useful. Geerts et al. (2021) found no evidence that a teleoperated robot facilitator increased brainstorming productivity compared with a human facilitator, while de Rooij et al. (2024) found that a robot facilitator's mood expressions affected collaboration, satisfaction, and co-creation performance. These results point to the same conclusion from different directions: the relevant object of study is not simply the presence of a robot, but the specific behaviour through which the robot intervenes in the group process.

This is especially important because creative ideation is phase-dependent. In divergent moments, a group seeks to open the idea space and generate alternatives, while in convergent moments, the same group compares, refines, or selects from ideas already raised (Isaksen et al., 2011). A robot question that is useful for generating more possibilities may be less useful when participants need to develop or evaluate an existing idea. The effect of a prompt, therefore, has to be interpreted relative to both its elicitation strategy and the phase in which it occurs.

The gap addressed in this study is the lack of prompt-level evidence on robot facilitation in creative group ideation. Existing work supports the idea that robots can shape group interaction, but it remains less clear how different spoken elicitation strategies shape the participant responses that immediately follow them. This study focuses on that designable unit of behaviour: the short robot prompt delivered at a specific point in the discussion.

1.2 Research goal and contribution

Based on this gap, the study investigates how Pepper, used as a social robot facilitator, elicits participant responses during a small-group ideation task. An elicitation prompt refers to Pepper's short spoken question or instruction intended to shape the next part of the human discussion. The main research question is:

> *How do robot elicitation strategies shape engagement and participant contribution substantiveness in creative group ideation processes?*

The study compares three verbal elicitation strategies: generative, elaborative, and perspective-shifting prompts. Each strategy is tested in both the divergent and convergent phases of the same ideation task. The study is exploratory and descriptive: it does not claim that one strategy is universally best, but examines whether different prompt types are associated with different patterns of engagement and substantive contribution.

The research question is divided into two subquestions:

> *SQ1: In each phase of the creative ideation process, how do elicitation strategies shape participant engagement?*

> *SQ2: In each phase of the creative ideation process, how do elicitation strategies shape participant contribution substantiveness?*

The main contribution is a prompt-level comparison of robot facilitation strategies in the same group ideation setup. By keeping the robot, task, session structure, and phase structure constant, the study focuses on the part of facilitation that can be directly designed: the spoken intervention the robot gives at a specific moment in the discussion. The next section grounds these strategies and the two outcome constructs in prior work before the methodology explains how they were operationalised in the experiment.

2 Related Work

The introduction motivates a prompt-level study of robot facilitation. This section grounds that choice in four parts: how robots can shape group interaction, why facilitation effects must be treated critically, how the three elicitation strategies relate to ideation theory, and how engagement and contribution substantiveness can be operationalised.

2.1 Robots in Group Interaction

Research on robots in groups shows that robots can shape interactions beyond the direct robot-user exchange. Sebo et al. (2020) review evidence that robots in groups and teams can affect coordination, communication, participation, and group dynamics. Gillet et al. (2024) make this focus explicit through interaction-shaping robotics, where robots are evaluated by how they influence interactions between other agents. This framing fits the present study because Pepper asks questions that may change how the human participants continue the discussion.

However, this literature does not by itself identify which verbal interventions are useful during ideation. A robot may shape interaction through embodiment, timing, gaze, affect, turn-taking, or speech content. The present study isolates one designable part of this broader space: the spoken elicitation strategy used at a scheduled moment in the discussion.

2.2 Robot Facilitation in Creative Tasks

Work on robot-supported creative facilitation shows both potential and limits. Geerts et al. (2021) found no evidence that brainstorming with a teleoperated robot facilitator produced higher productivity than brainstorming with a human facilitator, while de Rooij et al. (2024) found that a robot facilitator's mood expressions affected mood contagion, collaboration, process satisfaction, and co-creation performance. Together, these studies caution against treating robot facilitation as a uniform effect. The relevant question is not simply whether a robot is present, but which facilitator behaviours are used and when.

The present study therefore examines facilitation at a lower level of granularity. Instead of comparing robot and human facilitators or manipulating expressive style, it compares three kinds of robot questions within the same task, robot, and session structure. This cannot explain all effects of robot embodiment, but it can produce actionable evidence for prompt design.

2.3 Creative Ideation and Elicitation Prompts

Ideation tasks involve different conversational needs. Nijstad and Stroebe (2006) show that group ideation can be hindered by production blocking, while exposure to others' ideas can stimulate further idea search. Isaksen et al. (2011) distinguish divergent thinking, where possibilities are generated, from convergent thinking, where ideas are selected, evaluated, or developed. This distinction is a practical design choice rather than a claim that all creative work cleanly separates into two stages. It is useful here because it allows the same robot strategy to be compared under two different immediate task goals.

The elicitation strategies were derived from this literature rather than selected as arbitrary prompt types. First, divergent ideation depends on producing alternatives, so one strategy needs to invite further idea generation. This is reflected in the generative prompts, which ask participants for additional possibilities, adjacent options, or another way to approach the same problem. Their role is grounded in the idea-fluency side of creativity assessment, where the number of relevant ideas is treated as one important output of ideation (Kim, 2006), and in work on group idea generation showing that exposure to prior ideas can stimulate further search (Nijstad and Stroebe, 2006). The prompt wording follows Chin's productive-questioning move of *pumping*: short requests such as "What else?" that push for further responses, which is the basis recorded for these templates (Chin, 2007).

Second, creative contribution is not only about producing more ideas. Ideas also need to be clarified, specified, tested, or connected to practical constraints before they become usable. This motivates the elaborative prompts. Chin (2007) shows that productive teacher questioning, through moves such as requesting elaboration and posing reflective follow-up questions, can push learners to explain, justify, and develop their initial responses rather than only restate them. In this study, elaborative prompts translate that questioning function into the ideation task by asking participants how an idea would work, what evidence would support it, what could fail, or how it could be made more concrete. This links the strategy to elaboration as a creativity dimension (Kim, 2006) and to constructive or interactive contribution rather than passive agreement (Chi, 2009).

Third, ideation also benefits from reconsidering whether ideas are useful for someone other than the participants themselves. Grant and Berry (2011) found that perspective taking can support creativity by directing attention toward ideas that are useful to others, and Hoever et al. (2012) showed that perspective taking can help teams use diverse information for creative performance. This motivates the perspective-shifting prompts. The implementation is deliberately narrow: participants are not asked to generally become more empathetic, but to evaluate the current idea from one stable stakeholder perspective. In this study, all perspective-shifting prompts use first-year students as the target, because changing stakeholder roles across prompts would confound the strategy with the content of the perspective.

Together, the three strategies cover three complementary facilitation moves that are strongly connected to the two outcome constructs of this study: opening the idea space, developing the content already present, and reconsidering ideas through a user perspective. They are not presented as a complete set of creative facilitation prompts. They are a con-

trolled set of literature-grounded prompt types suitable for comparing how Pepper's spoken intervention shapes the response window that follows.

2.4 Measuring Engagement and Contribution Substantiveness

Engagement

Engagement is defined here as the process through which interaction partners establish, maintain, and end their perceived connection during a joint interaction (Sidner et al., 2005). In human-robot interaction, engagement is usually treated as a multi-component construct involving behavioural, social, and subjective signals rather than a single directly observable action (Sorrentino et al., 2024). For this reason, the present study measures engagement through response delay, speaking time, vocal activation, connection cues, and self-reported engagement.

While these measures do form a well-rounded set for a small, exploratory experiment such as this one, they do still exhibit meaningful drawbacks that must be mentioned. Fast responses can signal social connection in conversation (Templeton et al., 2022), but speed alone does not show whether participants are meaningfully engaged. Non-lexical cues such as backchannels, laughter, silence, and overlap can carry social meaning (Vinciarelli et al., 2015), but they also depend on speaking style, microphone quality and speech-to-text conversion. Literature-grounded instruments such as the UES and UES-SF are better suited for capturing participant self-reported engagement (O'Brien et al., 2018), but repeated full questionnaires would interrupt the session. The study, therefore, uses a short window-level engagement rating as a subjective complement to behavioural measures.

Contribution Substantiveness

Contribution substantiveness is defined as task-relevant talk that adds, develops, questions, compares, or builds on idea content beyond what has already been said. This definition separates substantive contribution from simply speaking more. Kim (2006) distinguishes idea fluency, the number of ideas produced, from elaboration, the amount of detail used to develop ideas. Chi (2009) similarly distinguishes passive engagement with information from more constructive and interactive forms of knowledge building. Barron (2003) adds that successful collaborative work depends on how participants take up and develop each other's proposals.

The present study therefore measures contribution substantiveness through three complementary indicators: distinct idea count, elaboration units, and consecutive turns on the same focal idea. Idea count captures breadth, elaboration units capture depth, and consecutive topic turns capture collaborative uptake. This combination is still a simplified operationalisation, but it avoids equating contribution quality with speaking duration alone. Together, the reviewed works motivate the methodological choice to compare elicitation strategies within the same task and phase structure, using separate measures for engagement and contribution substantiveness.

3 Methodology

3.1 Study Design

The study used a within-subjects design to compare how Pepper's elicitation strategy and the current ideation phase shaped participant responses while keeping the robot, task, technical setup, and session structure constant. The independent variables were elicitation strategy, with three levels: generative, elaborative, and perspective-shifting prompts, and ideation phase, with two levels: divergence and convergence. Each group received one prompt from each strategy in each phase, resulting in six elicitation windows per group. Figure 1 gives an at-a-glance overview of the full session, including the two phases, counterbalanced strategy order, elicitation windows, and context-only prompts that surround each scheduled elicitation.

The unit of analysis was the response window following a scheduled Pepper elicitation prompt. A window started when Pepper finished speaking. Under normal scheduling, it ended at the start of the next scheduled elicitation prompt, because the system delivered one elicitation prompt every four Pepper turns. This means that each window contains the elicitation prompt and the following three context-only Pepper prompts. The same cap was enforced for the last elicitation prompt of a phase or session: if participants chose to continue the discussion after the final scheduled strategy, any later context-only conversation was still excluded after the fourth Pepper-turn boundary. This rule kept the response windows comparable and prevented the final elicitation of a phase from receiving a longer measurement window simply because the group continued talking.

3.2 Participants and Task

The experiment included 40 participants arranged in 20 two-person groups. Participants were mainly TU Delft students recruited from the same student rowing context, so several pairs were familiar with each other before the session. This matched the informal student-group setting of the study, but is treated as a contextual limitation. Prior experience with Pepper, social robots, or generative AI tools was not used as an eligibility criterion and was not systematically measured.

Each session used the same task brief about improving the everyday campus experience for TU Delft students (Appendix A). The task was student-relevant, understandable without specialist knowledge, and open enough to allow multiple solution directions. Participants were informed in advance about the divergent and convergent phases and the scripted researcher interruptions for engagement ratings, phase transitions, and session ending. This minimised non-creative interruptions during ideation. All sessions were conducted in the same informal rowing-club setting and lasted approximately 30 minutes.

3.3 Experimental Procedure

Participants signed a validated consent form (Appendix B) and familiarised themselves with the task brief. The researcher started the audio, transcription, language-model, and Pepper speech pipeline, recorded an initial engagement rating, and the pair then began the task with Pepper. At the

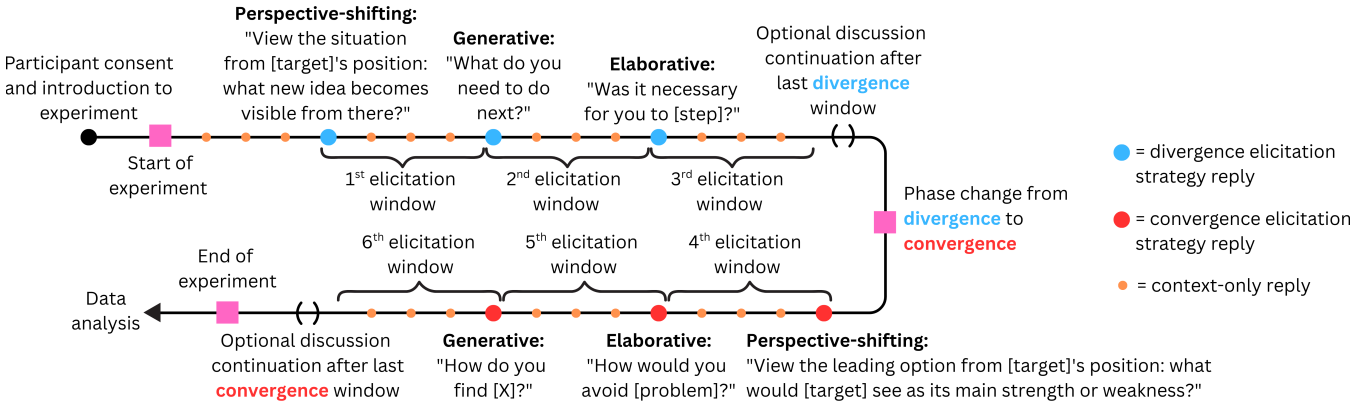

Figure 1: Overview of the experiment timeline. The session moves from consent and task familiarisation into divergence and convergence. Each phase contains three elicitation windows: one strategy prompt followed by up to three context-only replies within the four-Pepper-turn analysis window. Strategy order is counterbalanced, participants report engagement after each window, and continuation talk after a phase's final scheduled window is logged but excluded after the cap.

end of each elicitation window the researcher asked, "On a scale from 1 to 100, how would you rate your current engagement?". The researcher's speech was not analysed, and scores were captured by the participant microphones or stored as structured evaluation rows. Once the final divergence window was operationally complete, the researcher asked, "Divergence is complete. Continue or convergence?" and, on a positive answer, manually switched the phase with `CHANGE`. Convergence repeated the same three-window structure. After the final convergence window the researcher asked, "Continue?"; optional continuation was logged, but analysis still applied the four-Pepper-turn cap. Throughout, the researcher contributed no ideas, evaluations, or content steering and spoke only the scripted lines above.

3.4 System Pipeline

Pepper acted as a facilitator that delivered short spoken questions, not as an autonomous teammate generating ideas. The technical setup, therefore, had to support two linked functions: maintaining a live transcript of the participant discussion and generating a short Pepper utterance from the current context and, at scheduled points, the selected elicitation strategy.

Figure 2 shows one interaction cycle in the live system. Participant speech was captured through two external microphones, with one channel assigned to each participant; Pepper's built-in microphones were not used as the primary input because the measures required participant-level timing, speaking duration, overlap, and response behaviour. Audio was segmented using voice activity detection with a minimum speech duration of 0.35 seconds, a silence duration of 0.80 seconds for turn ending, and a maximum segment duration of 18 seconds. Each segment was sent to Deepgram Nova-3 through the EU endpoint for English speech-to-text transcription; filler words were requested, while punctuation and smart formatting were disabled so that the transcript stayed close to the live recogniser output. The returned text was stored with speaker label, timestamps, audio channel, source label, and audio-energy metadata.

When Pepper was triggered, the laptop sent the recent conversation context, phase, task theme, and prompt-bank state to a local `phi-3.5-mini` model served through LM Studio's OpenAI-compatible endpoint. Scheduled elicitation windows also included the selected strategy prompt; context-only turns used the same pipeline without one. The reply was sanitized, stripped of speaker labels and meta text, shortened to at most 24 spoken words, sent to Pepper's text-to-speech system, and logged with phase, strategy, prompt identifier, timestamps, and fallback status. The exact decoding settings and layered prompt structure are reported in Appendix C.

Because the sessions used non-native English speakers in an informal setting, the transcript contains periodic speech-to-text errors for short backchannels, laughter-like vocalisations, Dutch words, abbreviations, and uncommon names. Recognisable fillers and backchannels remained in the transcript, and short high-energy non-speech segments could be stored as acoustic nonverbal-event labels for the aggregate metrics. The transcript was therefore treated as a structured analysis record rather than as a perfect verbatim transcript.

The complete source code is available on GitHub at github.com/bogdanmicu12/Pepper, under the `elicitation` branch.

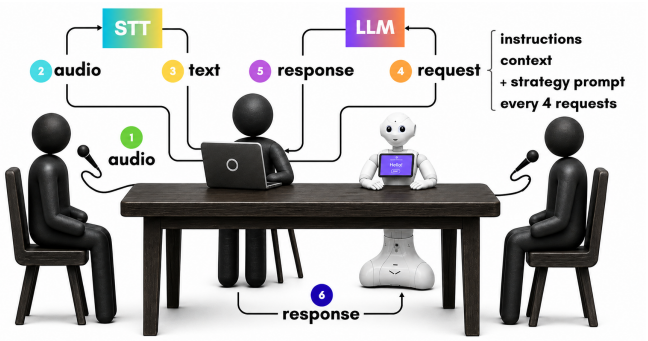

Figure 2: Overview of one live interaction cycle. Participant audio is transcribed, added to the laptop context, sent with instructions and the selected strategy prompt to the local model, and returned through Pepper's speech system.

3.5 Elicitation Strategy Enforcement

The elicitation strategies were controlled through the prompt bank in Appendix D, a strategy-order counterbalancing table, and a deterministic prompt selector. Generative templates asked for further possibilities or adjacent options, elaborative templates asked for explanation, evidence, failure conditions, or implementation detail, and perspective-shifting templates asked participants to reconsider the current idea from the perspective of first-year students. Each prompt-bank entry was assigned to one strategy, one ideation phase, and one literature basis, so the intervention was not improvised during the session.

The strategy-order table varied the order in which generative, elaborative, and perspective-shifting prompts appeared within each phase, so that one strategy was not always first, second, or third. At each scheduled elicitation point, the system selected the next unused strategy for the current phase and then selected a prompt template using the group identifier, theme identifier, phase, and strategy. The selected template was inserted into the LLM request together with the current task theme, phase, recent conversation history, latest participant turns, and last participant utterance. The model then turned the template into one short context-grounded question while preserving the intended strategy. The exact prompt identifier was stored in the transcript for every elicitation intervention, and all perspective-shifting prompts used the same stakeholder target, first-year students, to keep the condition stable across groups.

3.6 Measures

The measures operationalised engagement and contribution substantiveness. Because both constructs are multidimensional, no single metric was treated as sufficient. Engagement was grounded in Sidner et al.'s definition of how interaction partners establish, maintain, and end perceived connection (Sidner et al., 2005), and in HRI (Human-Robot Interaction) work that combines behavioural, vocal, social, and subjective indicators (Sorrentino et al., 2024). Engagement was therefore measured using response delay, speaking time, vocal activation, connection cue rate, and self-reported engagement.

Response delay captured immediate uptake of Pepper's elicitation prompt: the time between Pepper finishing and the first participant microphone/VAD speech onset in the response window. Participant start time came from microphone or VAD metadata, not Deepgram's returned transcript timestamp, because the latter corresponds to the completed audio segment. Fast responses can signal social connection and interactional alignment (Templeton et al., 2022), making response delay relevant but limited. Speaking time measured total participant speech duration inside the response window as a behavioural engagement indicator: longer talk can suggest sustained involvement, but it was interpreted cautiously because engagement concerns perceived connection and cannot be inferred from activity level alone (Sidner et al., 2005; Sorrentino et al., 2024).

The vocal activation score captured how active the participants sounded during the response window. It combined mean audio energy, speech rate, and long pauses into a standardised 0-100 score, following the broader use of vocal and behavioural signals in automatic engagement assessment (Sorrentino et al., 2024):

$$\begin{aligned} \mathrm{VAS}_w &= 50 + 50 \cdot \frac{A_w + B_w - C_w}{3}, \\ A_w &= z(\log(1 + \mathrm{RMS}_w)), \\ B_w &= z(\mathrm{words/s}_w), \\ C_w &= z(\mathrm{long\ pauses/min}_w). \end{aligned}$$

Long pauses were silence gaps longer than 3 seconds between participant turns, counting only the part beyond 3 seconds. The connection cue rate captured social responsiveness by counting backchannels, laughter events, and cooperative overlaps per participant speaking minute, because such nonlexical cues are socially meaningful in interaction analysis (Vinciarelli et al., 2015):

$$\mathrm{CCR}_w = \frac{\mathrm{backchannels}_w + \mathrm{laughter}_w + \mathrm{overlaps}_w}{T_w^{\mathrm{speech}}/60}.$$

Here T_w^{speech} is participant speaking time in seconds within window w.

Self-reported engagement was recorded as a 1-100 rating from each participant for the previous elicitation window. The two participant ratings were averaged for the window-level summaries. A short single-item rating was used instead of a repeated full questionnaire, such as the UES-SF (O'Brien et al., 2018), because the window-based design required six repeated measurements per group and a longer instrument would have interrupted the ideation flow.

Contribution substantiveness was measured separately from engagement, because active participation does not imply task-relevant idea development; idea quantity, elaboration, and interactive uptake are distinct aspects of contribution quality (Kim, 2006; Chi, 2009; Barron, 2003). *Distinct idea count* captured fluency by counting each new task-relevant proposal, feature, use case, problem framing, or evaluation criterion (Diehl and Stroebe, 1987; Nijstad and Stroebe, 2006; Kim, 2006). *Elaboration units* captured depth by counting concrete additions such as reasons, mechanisms, implementation steps, constraints, evidence checks, risks, mitigations, or expected outcomes (Kim, 2006; Chi, 2009). *Consecutive topic turns* captured collaborative uptake by counting the longest adjacent chain in which participants kept developing the same focal idea (Barron, 2003). The full coding rules are given in Appendix E.

3.7 Data Analysis

The analysis reconstructed elicitation windows from the structured transcript and computed all metrics per window before summarising them by phase and elicitation strategy. For each elicitation prompt, participant speech was assigned to the window according to the rules laid out in section 3.1. Behavioural and acoustic measures were computed automatically from timestamps, audio metadata, and transcript text, while contribution-substantiveness measures were coded manually with the same rules for every window. Because

the study is small and exploratory, the analysis reports group-level means with 95% confidence intervals rather than inferential tests. Thus, the goal is to identify patterns that may inform robot prompt design and later confirmatory work, not to establish definitive rankings between strategies.

4 Results

The analysis included 20 groups, each completing the divergent and convergent phases of the campus-experience ideation task and receiving one generative, one elaborative, and one perspective-shifting prompt per phase. Results are reported descriptively by phase and strategy as group-level means with 95% confidence intervals (n=20 groups per phase-strategy cell; intervals shown in Figures 3 and 4). Given the within-subjects structure and the exploratory sample size, no null-hypothesis significance tests were used to rank strategies. Instead, the analysis reports estimation descriptively through group-level means and 95% confidence intervals. The intervals are used to judge the stability and uncertainty of observed patterns, not as formal significance tests.

For SQ1, self-reported engagement showed the clearest pattern: convergence was higher than divergence for all three strategies, and this is the one engagement contrast whose confidence intervals are largely separated rather than overlapping (Figure 3). Convergence means were tightly grouped across generative, elaborative, and perspective-shifting prompts ($M = 65.7$, 65.5, and 65.6), while divergence means were lower ($M = 60.5$–61.0). The remaining engagement measures differed by strategy within a phase, but their confidence intervals overlap, so these contrasts are read only as directional. Response delay, measured from the end of Pepper's prompt to the first participant speech onset, was longest after divergence perspective-shifting prompts ($M = 8.3$ s) and shortest after convergence perspective-shifting prompts ($M = 4.7$ s). This measure should be interpreted as participant uptake latency rather than as an ordinary conversational turn gap, because it also includes the time needed for the two participants to orient to the robot's elicitation prompt and coordinate who would take the floor. Speaking time was highest in divergence after perspective-shifting and generative prompts; vocal activation was highest after convergence perspective-shifting prompts ($M = 53.4$); and connection cue rate was highest after divergence generative prompts ($M = 0.92$ cues per participant speaking minute).

For SQ2, the contribution measures show a clear phase distinction between opening and narrowing the solution space (Figure 4). Idea count was highest in divergence, especially after generative ($M = 7.45$) and perspective-shifting prompts ($M = 6.85$), and lower and more uniform across strategies in convergence ($M = 4.15$–4.45); the divergence–convergence gap is large relative to its confidence intervals, whereas the strategy differences within each phase are smaller and their intervals overlap. Elaboration units were highest after elaborative prompts in both phases (divergence $M = 14.85$, convergence $M = 14.65$), and consecutive same-subject turns followed the same pattern, peaking after convergence-elaborative prompts ($M = 12.65$). Descriptively, then, generative prompts were most associated with idea fluency and elaborative prompts with sustained, more developed discussion, though the within-phase strategy contrasts remain tentative given the overlapping intervals.

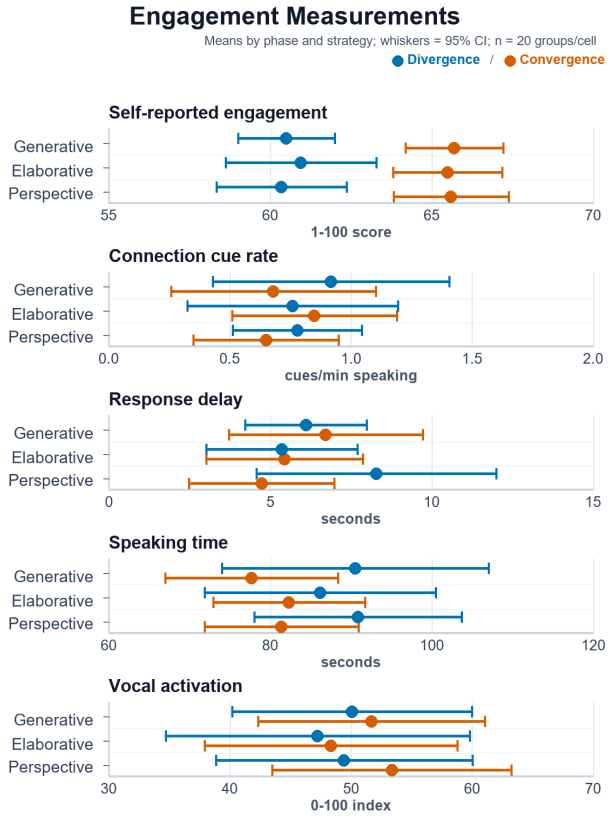

Figure 3: Engagement measurements by phase and elicitation strategy. Points show group means and whiskers show 95% confidence intervals.

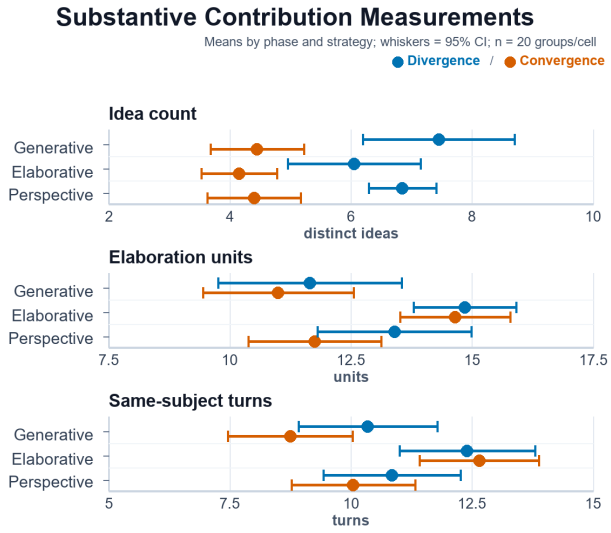

Figure 4: Substantive contribution measurements by phase and elicitation strategy. Points show group means and whiskers show 95% confidence intervals.

5 Discussion

The current results should not be read as a simple answer to whether Pepper "worked" as a facilitator or which elicitation strategy was best. The more constructive interpretation is that Pepper changed the conversation differently depending on ideation phase and question type. That is also the more interesting HRI conclusion: robot facilitation matters less as robot presence (Geerts et al., 2021) than as specific behaviour that reshapes how people respond to one another (Sebo et al., 2020; Gillet et al., 2024).

For SQ1, the clearest engagement pattern was the phase difference in self-reported engagement. Participants reported higher engagement in convergence than in divergence across all three strategies. One plausible reason is that convergence gave Pepper a more stable object to work with. In divergence, participants were still searching for what counted as a good campus-experience problem, so a prompt could feel like another interruption in an already open space. In convergence, concrete ideas were already on the table, and a robot question could help compare, justify, or narrow them. Engagement may therefore reflect perceived usefulness as well as liveliness: Pepper became easier to use once there was something definite to respond to. This extends work showing that robot facilitator behaviour can shape satisfaction and collaboration (de Rooij et al., 2024), but suggests that timing matters.

The response-delay pattern complicates this interpretation in an interesting way: the longest mean response delay occurred after divergence perspective-shifting prompts, while convergence perspective-shifting prompts produced the shortest mean delay. This suggests that perspective shifting depends strongly on timing. Perhaps asking participants to imagine a first-year student's perspective early in divergence may require an extra mental step before they have agreed what problem they are discussing. The same perspective in convergence may work more like an evaluation: would a first-year student understand this, trust it, or notice it? Perspective taking can support creativity (Grant and Berry, 2011; Hoever et al., 2012), but these results suggest that it needs something to attach to: without a shared idea it can create hesitation, with one it can sharpen judgment.

At the same time, response delay should be interpreted with the technical setup in mind. The measured delay starts when Pepper finishes speaking, but participants also experienced the earlier system delay while the laptop transcribed audio, updated context, queried the local model, cleaned the reply, and sent it through Pepper. That first delay is not the metric, but it matters socially. In most sessions, Pepper functioned less like a fluent conversational partner and more like a visible pause: the group talked, waited, listened, and then resumed. Since fast responses can signal social connection (Templeton et al., 2022), this helps explain why self-reported engagement stayed moderate even when speaking time and vocal activation were high. The system could still stimulate useful discussion, but the slow pipeline likely made Pepper feel less naturally present than a human facilitator.

The other engagement measures show why engagement should remain multi-component (Sidner et al., 2005; Sorrentino et al., 2024). Divergence generative prompts produced the highest connection cue rate, which fits the rhythm of brainstorming: asking what else belongs on the board invites quick agreements, corrections, and cooperative uptake. By contrast, vocal activation was highest after convergence perspective-shifting prompts, suggesting that groups became more animated when judging ideas from a concrete user perspective. Speaking time was highest in divergence after perspective-shifting and generative prompts, but longer talk did not necessarily mean stronger self-reported engagement or more substantive contribution. Backchannels, laughter, silence, and overlap can carry social meaning (Vinciarelli et al., 2015); here, they helped separate uptake quickness, vocal energy, sustained speaking, and perceived engagement.

For SQ2, the contribution measures mapped more directly onto the intended prompt strategies. Generative prompts were most associated with idea count, especially in divergence. This was expected, but still valuable as a manipulation check: generative prompts kept participants in open search mode rather than nudging them toward early evaluation (Nijstad and Stroebe, 2006; Isaksen et al., 2011). Elaborative prompts showed a different effect, but just as unsurprising. They produced the highest elaboration-unit counts and longest same-subject turn chains in both phases, suggesting that they encouraged participants to stay with an idea long enough to add evidence, risks, implementation details, or constraints. This matches productive-questioning and collaborative-uptake work (Chin, 2007; Barron, 2003), and supports measuring contribution substantiveness separately from speaking time: a group can talk a lot without developing an idea, or develop one in a compact exchange (Kim, 2006; Chi, 2009).

This distinction is also visible in the conversation content. The experiments coincided with student-council elections, and this participant cluster included socially active students already exposed to campus debates. As a result, stronger, more critical, and more socially aware discussions emerged. Several groups did not only list problems, but also asked what realistic sources of evidence would be. Quiet recovery and contemplation rooms, for example, felt credible partly because Lijst Beta had mentioned similar needs. Some groups even treated student-council turnout itself as a campus-experience issue, proposing a "what changed because of your vote" panel. These examples matter because the task became not only about original campus improvements, but also about using social context to decide which problems felt legitimate.

I would argue this contextual awareness acted as both a strength and a caution. It made the discussions richer: participants did not stay with generic complaints about apps or buildings, but connected ideas to actual campus debates such as police data disclosure, tram-route confusion, student housing pressure, and support-service visibility. At the same time, the content cannot be treated as neutral evidence of what all TU Delft students would prioritise. The robot elicitation strategy shaped the form of the discussion, but the participants' shared environment shaped the material they had available. This is acceptable for the research question, because the analysis compares how prompt strategies shaped response windows, not which campus problems matter most. Pepper's

prompts should therefore be understood as reorganising and extending the knowledge participants brought into the room, rather than producing a neutral map of student concerns.

Overall, the results support the utility of observing robot facilitation at the prompt level. Pepper did not need to be the creative agent in the room. Its useful role was to ask questions that changed what participants did next: broaden possibilities, deepen one option, or reframe an idea through a stakeholder. The findings did not show one strategy dominating engagement or contribution substantiveness across phases. What they did reveal are underlying trends on more granular levels: generative prompts keep the floor open for longer, elaborative prompts create depth, and perspective-shifting prompts become powerful once there is already an idea to evaluate. Designing robot facilitators should therefore focus less on making the robot generally "more engaging" and more on choosing the right question at the right moment.

5.1 Limitations

Several limitations should guide interpretation. First, the study is exploratory and based on a small current dataset. Confidence intervals overlap in several conditions, so the results are descriptive patterns rather than proof that one strategy is statistically superior. Second, the participant sample was specific: TU Delft students recruited mainly through an informal rowing-club context, with many socially active participants during the student-council election period. This made the conversations grounded, but limits generalisation to less involved students or to groups outside TU Delft.

Third, the technical pipeline shaped the interaction. Response delay was measured from the end of Pepper's spoken prompt to the first participant speech onset, but participants also experienced the unmeasured delay before Pepper spoke. Speech-to-text errors affected the transcript and sometimes Pepper's next response, especially for local names, abbreviations, short backchannels, Dutch words, and homophones such as "to", "too", and "two". These errors are realistic for the setup, but they mean the transcript should be treated as a structured analysis record rather than as a perfect account of what was said.

Finally, the measures focus on immediate response windows after elicitation prompts. This design is appropriate for comparing prompt-level effects, but it does not directly evaluate final solution quality and does not capture every useful exchange that happened after the third and sixth windows. A strong final direction might emerge from several small exchanges that no single response window fully contains.

5.2 Future Work

Future work should first separate the effect of prompt strategy from the effect of system timing. A faster or streamed pipeline, with less visible delay across transcription, language-model generation, and speech, could change how participants experience Pepper's interventions.

A second direction is to test the elicitation strategies in a between-subjects design if a sufficiently large participant pool is available. The present within-subjects structure, where every group experienced all three strategies in both phases, made comparison feasible with limited groups, but later prompts may have been shaped by earlier prompts, learning effects, and ideas already produced in previous windows. Assigning groups to different strategy conditions would reduce carry-over and make it easier to test whether the observed patterns came from the prompt strategy itself rather than its position in an ongoing discussion.

Finally, future work should connect prompt-level process measures to final creative outcomes. External judges could rate final solutions on originality, usefulness, feasibility, or clarity, adding a more content-oriented measure of substantive contribution. This would ask not only whether Pepper changed the discussion, but whether those changes improved the ideas groups carried forward. The broader opportunity is to design social robots that support human creativity not by taking over the creative work, but by asking the question that helps a group notice the next useful move.

6 Conclusion

This study examined how Pepper's spoken elicitation strategies shape engagement and participant contribution substantiveness during creative group ideation. The central finding is that robot facilitation is not best understood as a general property of having a robot in the room. Its effect depends on the question Pepper asks, the ideation phase in which it asks it, and the aspect of the discussion being measured. Pepper functioned as a phase-sensitive prompting system: it did not provide ideas itself, but tried to influence the kind of thinking participants did next.

The results answer the research question in a phase-sensitive way. Generative prompts were most closely associated with idea breadth, especially during divergence, where keeping the search space open is the main task. Elaborative prompts were most closely associated with contribution substantiveness, producing more elaboration units and longer same-subject discussion. Perspective-shifting prompts appeared most useful once participants already had ideas to evaluate, especially in convergence, where the first-year-student perspective gave groups a concrete lens for judging whether an option would be understandable, relevant, and believable. Engagement also proved multi-dimensional: speaking time, uptake speed, vocal activation, connection cues, and self-reported engagement did not always move together. Because the study is exploratory and confidence intervals overlap across several measures, these findings should be treated as design evidence rather than definitive strategy rankings.

The main contribution is therefore a prompt-level account of social robot facilitation in creative ideation. By holding the task, robot, phase structure, and scheduling rule constant, the study isolated a design choice future systems can act on: which question should the robot ask at a given moment? Effective robot facilitation should match the prompt to the group's conversational need. When the group needs more possibilities, the robot should open the space; when ideas are vague, it should ask for evidence, constraints, and mechanisms; when ideas are ready to be judged, it can introduce a stable stakeholder perspective.

7 Responsible Research

The study involved a social robot, participant speech, transcript data, manual coding, and automated analysis scripts. Responsible research was therefore treated as both an ethical issue and a reproducibility issue. The work was guided by the principles of honesty, scrupulousness, transparency, independence, and responsibility described in the Netherlands Code of Conduct for Research Integrity and by TU Delft's emphasis on verifiable research practice and careful handling of data and information (net, 2018; Delft University of Technology, 2018). These principles were applied by documenting the experimental procedure, logging robot prompts and timestamps, separating raw participant material from analysis outputs, and reporting descriptive patterns without presenting them as strong causal evidence.

7.1 Data, Privacy, and Participant Protection

Participants received an informed-consent form before participation and were told that participation was voluntary, that they could withdraw at any time, and that their data could be removed after participation on request. The experiment required participant speech to be recorded and transcribed in order to reconstruct response windows after Pepper's prompts. This creates a privacy risk because spoken interaction can contain identifying details, personal phrasing, or contextual information even when names are not intentionally collected.

For this reason, raw audio and identifiable transcripts are not suitable for unrestricted public release. Audio was used during the experiment for real-time speech-to-text conversion and then deleted. Deepgram was used as a third-party speech-to-text API, so the study used the EU endpoint to reduce data-transfer risk and treated the API step as an explicit privacy consideration rather than as a neutral technical detail. Reproducibility is instead supported through non-identifying research materials: the prompt bank, counterbalancing table, analysis scripts, data schemas, manual coding rules, and aggregated outputs.

The study also used data minimisation. The task did not require names, student numbers, grades, health information, or political preferences, and participants were asked to avoid local abbreviations. Some campus-related discussion still occurred during the student council election period, so political or organisational references could appear naturally in the conversation. These references were treated as contextual discussion material, not as individual political data. The analysis was conducted at group-window level, and examples were reported only when they helped explain the interpretation of the findings.

Pepper was used as a facilitator that elicited participant contributions rather than as a teammate that generated ideas on behalf of the group. This matters ethically because the system should not obscure whether an idea came from a participant or from the technology. Participant contributions were analysed at group-window level rather than used to evaluate individual ability, and self-reported engagement scores were treated as subjective research measures, not as performance judgements. The researcher also avoided giving content ideas during the sessions, because doing so would have blurred the boundary between the participant discussion, the robot intervention, and the researcher's own preferences.

7.2 Reproducibility of the Method

Reproducibility concerns whether consistent results can be obtained using the same input data, code, computational steps, and analysis conditions, whereas replicability concerns whether a new study addressing the same research question obtains consistent findings with newly collected data (National Academies of Sciences, Engineering, and Medicine, 2019). The main reproducibility target of this study was the analysis pipeline: given the same transcript, intervention log, manual coding file, and configuration, another researcher should be able to reconstruct the same elicitation windows, compute the same measures, and regenerate the same summary tables and figures.

To support this, the project stores the experimental and analysis materials in explicit files. The prompt-bank templates are stored in `llm/prompts/prompt_bank.csv`, the strategy order is stored in `llm/design/counterbalancing_elicitation.csv`, and the live system logs robot and participant events to structured CSV files. The analysis script reconstructs elicitation windows, computes behavioural and acoustic measures, merges manually coded contribution-substantiveness measures, and writes CSV summaries and figures

At the same time, the live human-robot interaction cannot be made perfectly reproducible in a deterministic sense. Pepper's behaviour depends on hardware availability, microphone placement, speech recognition quality, the local LM Studio model, and the participants' responses in the moment. The language-model step is also not fully transparent in the way a hand-written script would be, even though the model, endpoint, prompt layers, stop sequences, and fallback handling were documented. For this reason, the study logs the actual prompt identifiers, selected strategies, phases, timestamps, and robot utterances. These logs make the analysis reproducible and make the experimental procedure replicable, even if the social interaction itself cannot be replayed exactly.

The manual contribution measures introduce another reproducibility limitation. Distinct ideas, elaboration units, and consecutive topic turns require judgement, and this study did not include a second independent coder or an inter-rater reliability estimate. This is reported in the study limitations subsection rather than hidden. The coding rules were made explicit in the appendix to make the decisions auditable and to allow later work to repeat the coding with additional coders.

7.3 Bias, Uncertainty, and Broader Impact

Several sources of bias remain. The participant sample and task context limit generalisation, automatic transcription can introduce errors, and the local language model may reflect biases in the data and instruction-following behaviour on which it was trained. The task was also campus-specific, and many participants were socially active students, so their priorities may not represent quieter, less involved, international, commuting, disabled, or first-year students equally well. The study therefore reports descriptive patterns and uncertainty rather than overstating the strength of the results.

Robot-supported ideation could help groups continue discussions, consider neglected perspectives, or develop ideas more concretely. The same technology could also create pressure to respond to a machine, overstate the authority of automated prompts, or normalise detailed monitoring of group interaction. The design of this study keeps Pepper's role narrow: it asks short elicitation questions, does not score participants during the session, and does not replace human ownership of the ideas. This boundary is important for future use: a robot facilitator should support human creativity and reflection, not become a hidden evaluator of students or a tool for extracting behavioural data without a clear research purpose.

References

(2018). Netherlands code of conduct for research integrity.

Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences*, 12(3):307–359.

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, 1(1):73–105.

Chin, C. (2007). Teacher questioning in science classrooms: Approaches that stimulate productive thinking. *Journal of Research in Science Teaching*, 44(6):815–843.

de Rooij, A., van den Broek, S., Bouw, M., and de Wit, J. (2024). Co-creating with a robot facilitator: Robot expressions cause mood contagion enhancing collaboration, satisfaction, and performance. *International Journal of Social Robotics*, 16:2133–2152.

Delft University of Technology (2018). TU Delft Vision on Integrity 2018–2024.

Diehl, M. and Stroebe, W. (1987). Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology*, 53(3):497–509.

Geerts, J., de Wit, J., and de Rooij, A. (2021). Brainstorming with a social robot facilitator: Better than human facilitation due to reduced evaluation apprehension? *Frontiers in Robotics and AI*, 8:657291.

Gillet, S., Vázquez, M., Andrist, S., Leite, I., and Sebo, S. (2024). Interaction-shaping robotics: Robots that influence interactions between other agents. *ACM Transactions on Human-Robot Interaction*, 13(1):12:1–12:23.

Grant, A. M. and Berry, J. W. (2011). The necessity of others is the mother of invention: Intrinsic and prosocial motivations, perspective taking, and creativity. *Academy of Management Journal*, 54(1):73–96.

Hoever, I. J., van Knippenberg, D., van Ginkel, W. P., and Barkema, H. G. (2012). Fostering team creativity: Perspective taking as key to unlocking diversity's potential. *Journal of Applied Psychology*, 97(5):982–996.

Isaksen, S. G., Dorval, K. B., and Treffinger, D. J. (2011). *Creative Approaches to Problem Solving: A Framework for Innovation and Change*. SAGE Publications, Los Angeles, 3 edition.

Kim, K. H. (2006). Can we trust creativity tests? a review of the torrance tests of creative thinking. *Creativity Research Journal*, 18(1):3–14.

National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. National Academies Press.

Nijstad, B. A. and Stroebe, W. (2006). How the group affects the mind: A cognitive model of idea generation in groups. *Personality and Social Psychology Review*, 10(3):186–213.

O'Brien, H. L., Cairns, P., and Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies*, 112:28–39.

Sebo, S., Stoll, B., Scassellati, B., and Jung, M. F. (2020). Robots in groups and teams: A literature review. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):176:1–176:36.

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., and Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1–2):140–164.

Sorrentino, A., Fiorini, L., and Cavallo, F. (2024). From the definition to the automatic assessment of engagement in human–robot interaction: A systematic review. *International Journal of Social Robotics*, 16:1641–1663.

Templeton, E. M., Chang, L. J., Reynolds, E. A., Cone LeBeaumont, M. D., and Wheatley, T. (2022). Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences*, 119(4):e2116915119.

Vinciarelli, A., Chatziioannou, P., and Esposito, A. (2015). When the words are not everything: The use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls. *Frontiers in ICT*, 2:4.

A Task brief

In this session, you will discuss ideas for improving the everyday campus experience for TU Delft students. This can include any part of student life on or around campus, such as finding information, moving between buildings, planning study or social activities, food and drinks, events, group work, facilities, communication, wellbeing, accessibility, or other practical issues students encounter during a normal day.

Please speak in English. To make the transcript understandable, avoid using local names, abbreviations, or shorthand without explaining them. For example, say "the rowing club" instead of "Proteus", "the campus store" instead of "Spar", "Policy Management building" instead of "TPM", or "Industrial Engineering building" instead of "IO". If you do use a specific name or abbreviation, please briefly explain what it means.

During the session, Pepper may respond automatically or after you address it directly. If you want to speak to Pepper directly, first address it with "Pepper" and then continue with what you want to ask or say. Please be patient after addressing

Pepper, because it may take a few seconds for the system to detect the speech, transcribe it, generate a response, and make Pepper speak. Avoid calling Pepper repeatedly while waiting for a response, as this can create duplicate triggers or interrupt the current interaction.

The session will last approximately 30 minutes and has two phases. First, you will do a divergent phase. In this phase, the goal is to generate possible ideas, explore different directions, mention rough ideas even if they are not fully developed yet, and build on each other's suggestions. Then, the convergence phase will follow. In this phase, the goal is to make promising ideas more concrete, compare alternatives, and move toward one or more final solution directions.

I will only interrupt at specific moments. After each measured part of the discussion, I will ask each of you: "On a scale from 1 to 100, how would you rate your current engagement?" Please, each say one number from 1 to 100 that reflects how engaged you personally felt during that part. After the divergent phase is complete, I will ask: "Divergence is complete. Continue or convergence?" If you are ready to move on, I will switch the system to the convergent phase. If you want to continue generating ideas, you can continue the discussion and let me know when you are ready to change to the second phase. After the final measured part of the convergent phase, I will ask: "Continue?" If you want to continue, you can keep discussing with Pepper using normal follow-up questions and decide for yourself when the discussion is finished.

B Consent Form

INFORMED CONSENT FORM

You are being invited to participate in a research study about creative collaboration with one or more social robots. This study is being done by Ruben Weijers and Catharine Oertel from the TU Delft.

The purpose of this study is to understand the effect of a social robot's interaction style on human collaboration. The session will take approximately 30-50 minutes. The data will be used for BSc theses and potential publication. You will be asked to complete a brief questionnaire, collaborate with a human partner and a social robot on an open-ended challenge, complete further questionnaires, and take part in a short group interview with your partner about your experience. During the session, we will collect: (1) audio and/or video recordings of the session, (2) your responses to questionnaires, and (4) basic demographic information (such as age, gender, and country of origin) used only to describe the overall participant sample.

To the best of our ability, your answers in this study will remain confidential. We will minimize any risk by removing any mention of names or sensitive information from data.

Your participation is entirely voluntary and you may withdraw at any time during the session without giving any reason. During the session, you are free to stop at any time without providing a reason, and you are free to request the deletion of your data. You will not be financially compensated for your time.

For questions or requests to delete your data

PLEASE TICK THE APPROPRIATE BOXES

	Yes	No
A: GENERAL AGREEMENT – RESEARCH GOALS, PARTICIPANT TASKS AND VOLUNTARY PARTICIPATION		
I have read and understood the above information.	☐	☐

I have been able to ask questions about the study and my questions have been answered to my satisfaction.	☐	☐
I consent voluntarily to be a participant in this study and understand that I can withdraw from the study at any time, without having to give a reason.	☐	☐
I understand that taking part in the study involves discussion with a conversational robot.	☐	☐
I understand that taking part in the study involves completing questionnaires and a short group interview about my experience.	☐	☐
I understand that the interview takes place with my partner present, and that I should not share anything I would not want my partner to hear.	☐	☐
I understand that the study will last approximately 45 minutes.	☐	☐
I understand that the session will be audio and video recorded	☐	☐
B: POTENTIAL RISKS OF PARTICIPATING (INCLUDING DATA PROTECTION)		
I understand that my data will be treated confidentially, that any direct identifiers (such as my name) will be replaced by a pseudonym for analysis, and that names mentioned during the session will be removed from transcripts.	☐	☐

I understand that I may request deletion of my data up until June 15th, after which deletion may no longer be possible	☐	☐
I understand that I must not provide any personally identifiable information such as phone number, email address or password. If I do this, it will be removed from the recordings and this may destroy the consistency of the data.	☐	☐
I understand that anonymised research data will be stored for 10 years in accordance with TU Delft's Research Data Framework Policy.	☐	☐
I understand that anonymised outputs from this study, including redacted transcripts, coded behavioural data, aggregated survey responses, and individually screened quotes, may be shared with other researchers on request, under a Creative Commons Attribution (CC BY 4.0) licence requiring attribution to the original researchers. I understand that raw audio and video recordings will not be shared outside the research team.	☐	☐
C: RESEARCH PUBLICATION, DISSEMINATION AND APPLICATION		
I understand that after the research study the de-identified information I provide will be used for BSc theses / potential publications.	☐	☐
I agree that my responses can be quoted anonymously in research outputs.	☐	☐

I agree that some parts of the conversation and task outcome can be shown in research outputs (BSc theses, potential publications) or snapshots of them can appear anonymously.	☐	☐
D: (LONGTERM) DATA STORAGE, ACCESS AND REUSE		
I give permission for the anonymised data that I provide to be archived in the 4TU repository so it can be used for future research and learning.	☐	☐
I understand that access to this repository is restricted and that other researchers may request access for non-commercial research and teaching purposes.	☐	☐

Signatures

_____ _____ _____ Name of participant
Signature Date

I, as researcher, have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

_____ _____ _____
Researcher name Signature Date

Study contact details for further information: Ruben Weijers

C LLM and Logging Infrastructure

The live system used a layered prompting pipeline. The first layer was a base instruction defining Pepper as a short spoken facilitator in a two-person brainstorming session. The exact base instruction was: "You are Pepper in a two-person brainstorming session. Respond naturally and conversationally, as if speaking aloud. Stay grounded in the latest participant turns and the recent history. Ask one concise, context-grounded question that nudges a more creative angle. Avoid summarizing, deciding, or proposing a full solution unless directly asked. Do not sound like a workshop facilitator: do not begin with Considering, Given, In light of, or Based on. Do not mention engagement scores, metrics, or the experiment unless the participants directly ask about them. Do not write dialogue labels, notes, markdown, analysis, explanations of your behavior, or multiple possible replies. If participants are joking, testing, or using profanity, stay calm and redirect briefly without scolding."

The second layer added the active elicitation mode. For Perspective-shifting, the exact mode guidance was: "Use perspective-taking to broaden idea search before judgement." For Generative prompting, it was: "Use creative brainstorming techniques to generate new ideas." For Elaborative prompting, it was: "Push abstract ideas into concrete, testable details and evidence checks."

The third layer added the current task context. The exact context template was: "Theme: [theme]. Phase: [phase]. Recent conversation history: [recent conversation history]. Latest uninterrupted participant turns: [latest participant turns]. Last participant utterance: [last participant utterance]. Seed ideas: [seed ideas]." These fields were filled automatically from the running session.

The fourth layer added the selected counterbalanced prompt-bank template. The exact prompt template was selected from the prompt bank reported in Appendix D. It was inserted using the following instruction: "Intervention prompt to adapt briefly to this moment: [selected prompt-bank template]. Use what the participants just said to ask the intervention as a natural creativity-boosting question."

The final layer was a reply contract. The exact reply contract was: "Output only Pepper's spoken words. Prioritize the latest participant turns over older history; do not invent a new topic. Use one short creativity-boosting question whenever possible, preferably starting with What if, What would, Which small, or How might. Do not summarize the discussion before asking. No speaker labels such as Pepper: or Robot:. No notes, markdown, separators, lists, or meta-commentary. Maximum 24 words."

The local language model was called through LM Studio using the endpoint `http://127.0.0.1:1234/v1/chat/completions`. The implementation used `phi-3.5-mini-3.8b-instruct`, temperature `0.35`, and `max_tokens = 600`. Stop sequences were used to prevent the model from producing speaker labels, notes, or extra dialogue turns. The exact stop sequences were `\nParticipant`, `\nParticipant 1`, `\nParticipant 2`, `\nRobot:`, `\nPepper:`, `\n---`, `\nNote:`, and `\n**Note`. After generation, the response was sanitized again before being sent to Pepper.

The system also included placeholder handling. Perspective-shifting templates used the placeholder `[target]`, which was replaced with "a first-year student". Other placeholders, such as `[A]`, `[B]`, `[output]`, and `[problem]`, were treated as abstract template markers. The exact placeholder instruction was: "If the intervention prompt contains bracketed placeholders such as [A], [B], [X], [goal], [output], [values], [result], [step], [problem], or [target], replace them silently with relevant content from the recent conversation before speaking. Never say bracketed placeholders aloud."

The transcript stored one row per relevant event. Participant rows stored session identifier, group identifier, timestamp, start and end time, speaker label, text, phase, audio mode, audio device, audio channel, mean audio energy, trigger words, overlap status, and acoustic confidence. Robot rows stored the robot reply, phase, strategy, prompt identifier, source of the reply, and fallback reason if applicable. This structure made it possible to reconstruct response windows and compute engagement and contribution measures at the elicitation-prompt level.

D Prompt Bank

ID	Strategy	Phase	Basis	Template
P-D-01	Perspective-shifting	Divergence	(Hoever et al., 2012)	Try to take [target]'s perspective as much as possible: what might [target] be noticing that we have not considered yet?
P-D-02	Perspective-shifting	Divergence	(Hoever et al., 2012)	View the situation from [target]'s position: what new idea becomes visible from there?
P-D-03	Perspective-shifting	Divergence	(Hoever et al., 2012)	Try to understand how [target] views this situation: what concern or opportunity might shape [target]'s view?
P-D-04	Perspective-shifting	Divergence	(Hoever et al., 2012)	Ask what is important to [target] then suggest one idea that could respond to that priority.
P-C-01	Perspective-shifting	Convergence	(Hoever et al., 2012)	As you narrow the options try to take [target]'s perspective: which option would still make sense from [target]'s point of view?
P-C-02	Perspective-shifting	Convergence	(Hoever et al., 2012)	View the leading option from [target]'s position: what would [target] see as its main strength or weakness?
P-C-03	Perspective-shifting	Convergence	(Hoever et al., 2012)	Try to understand how [target] views this choice: what concern should be discussed before deciding?
P-C-04	Perspective-shifting	Convergence	(Hoever et al., 2012)	Ask what is important to [target] then adjust the option so it better integrates that priority.
P-D-05	Generative	Divergence	(Chin, 2007)	What else?
P-D-06	Generative	Divergence	(Chin, 2007)	How is [target] going to solve the problem?
P-D-07	Generative	Divergence	(Chin, 2007)	In order to find [goal], what must we do?
P-D-08	Generative	Divergence	(Chin, 2007)	What do you need to do next?
P-C-05	Generative	Convergence	(Chin, 2007)	How do we find [output]?
P-C-06	Generative	Convergence	(Chin, 2007)	With these [values], how do we find [result]?
P-C-07	Generative	Convergence	(Chin, 2007)	What do we do next?
P-C-08	Generative	Convergence	(Chin, 2007)	How do you find [X]?
P-D-09	Elaborative	Divergence	(Chin, 2007)	What is the difference between [A] and [B]?
P-D-10	Elaborative	Divergence	(Chin, 2007)	Where does [X] come into the calculation?
P-D-11	Elaborative	Divergence	(Chin, 2007)	Was it necessary for you to [step]?
P-D-12	Elaborative	Divergence	(Chin, 2007)	Could you have done with fewer steps?
P-C-09	Elaborative	Convergence	(Chin, 2007)	Can you modify it in such a way that you do not need to do unnecessary steps?
P-C-10	Elaborative	Convergence	(Chin, 2007)	What do you think are some of the problems and how do you think you can overcome it?
P-C-11	Elaborative	Convergence	(Chin, 2007)	How would that affect the accuracy?
P-C-12	Elaborative	Convergence	(Chin, 2007)	How would you avoid [problem]?

Table 1: Prompt-bank templates implemented in the system. Prompt templates were counterbalanced within each strategy-phase condition, and the exact selected template was logged through its prompt identifier.

E Manual Coding Labels for Contribution Substantiveness

This appendix defines the individual coding labels used when manually identifying distinct ideas, elaboration units, and excluded transcript material. The purpose of these definitions was to make the coding conservative and consistent across elicitation windows.

- **Task-relevant proposal.** A task-relevant proposal was counted when a participant introduced a possible solution, action, intervention, design direction, or recommendation that directly addressed the ideation task. It was enforced as a new idea only if it added a new direction to the discussion, rather than restating an existing one. This follows idea-fluency coding in creativity assessment, where relevant generated ideas are counted while repetitions are excluded (Kim, 2006).
- **Feature.** A feature was counted when a participant proposed a specific attribute, function, component, or capability of a solution. Features were counted as new ideas only when they changed what the proposed solution would include or do. Features that merely clarified an already stated solution without adding new content were treated as elaboration instead. This distinction follows the separation between idea fluency and elaboration: fluency concerns new idea content, while elaboration concerns further specification (Kim, 2006).
- **Use case.** A use case was counted when a participant described a concrete situation, user group, context, or scenario in which an idea could be applied. A use case was counted as a new idea only when it introduced a materially different application direction. If it only illustrated an already stated proposal, it was counted as an elaboration unit. This rule reflects the role of examples and contextualization in developing ideas beyond their initial statement (Kim, 2006; Chi, 2009).
- **Problem framing.** A problem framing was counted when a participant redefined what problem should be solved, who the relevant stakeholder was, what the main need was, or what aspect of the task mattered most. It was counted as a new idea when it redirected the discussion toward a different interpretation of the task. This is grounded in creative problem-solving work that treats problem understanding and reframing as part of the ideation process, not only solution generation (Isaksen et al., 2011).
- **Evaluation criterion.** An evaluation criterion was counted when a participant introduced a standard by which ideas could be judged, such as feasibility, cost, inclusiveness, ease of adoption, safety, novelty, or usefulness for a stakeholder. It was counted as a distinct idea when it added a new basis for comparing or selecting ideas. This aligns with the distinction between divergent and convergent thinking, where convergence involves evaluating, comparing, and selecting ideas (Isaksen et al., 2011).
- **Rewording.** Rewording referred to saying the same idea in different words without adding a new function, stakeholder, reason, constraint, or implication. Rewording was not counted as a new idea or elaboration. This exclusion prevented fluency scores from being inflated by surface variation rather than substantive content (Kim, 2006).
- **Agreement.** Agreement referred to short acceptance or endorsement of another participant's contribution, such as saying that an idea was good, useful, or acceptable. Agreement was not counted unless it included a reason, consequence, modification, or additional detail. This follows the distinction between passive participation and constructive contribution (Chi, 2009).
- **Example of the same proposal.** An example of the same proposal was not counted as a new idea when it only illustrated an already stated idea. It was counted as an elaboration unit if the example made the idea more concrete, showed how it would appear in practice, or added task-relevant detail. This rule preserves the distinction between generating more ideas and developing existing ideas (Kim, 2006).
- **Reason.** A reason was counted as an elaboration unit when a participant explained why an idea might work, why it mattered, or why it should be preferred. Reasons were not counted when they were generic evaluations without explanation. This reflects constructive contribution, where participants add explanatory content rather than merely reacting (Chi, 2009).
- **Mechanism.** A mechanism was counted as an elaboration unit when a participant explained how an idea would produce its intended effect. This included causal explanations, interaction flows, or process descriptions. Mechanisms were treated as elaboration because they develop an idea beyond its initial proposal (Kim, 2006; Chi, 2009).
- **Implementation step.** An implementation step was counted as an elaboration unit when a participant described a concrete action needed to carry out an idea. Examples include assigning a role, introducing a tool, changing a process, or sequencing actions. This was coded as elaboration because it specifies how an idea could be realized in practice (Kim, 2006).
- **Constraint.** A constraint was counted as an elaboration unit when a participant identified a limitation, requirement, dependency, or boundary condition affecting an idea. Constraints included time, resources, user needs, technical limitations, or institutional rules. Constraints were included because they make ideas more specific and support convergent development (Isaksen et al., 2011).
- **Evidence check.** An evidence check was counted as an elaboration unit when a participant questioned whether an idea was supported, realistic, testable, or consistent with known information. It was not counted when the participant only expressed doubt without specifying what needed to be checked. This was grounded in the

role of questioning and explanation in productive discussion (Chin, 2007; Chi, 2009).

- **Risk.** A risk was counted as an elaboration unit when a participant identified a possible negative outcome, failure mode, or unintended consequence of an idea. Risks were counted only when they referred to a specific issue, not when they expressed general dislike. This supports substantive idea evaluation during convergence (Isaksen et al., 2011).
- **Mitigation.** A mitigation was counted as an elaboration unit when a participant proposed a way to reduce, avoid, or respond to a risk or constraint. It was coded separately from the risk if it added a distinct solution-oriented detail. This reflects constructive development because the participant modifies or strengthens an existing idea (Chi, 2009).
- **Expected outcome.** An expected outcome was counted as an elaboration unit when a participant described what would likely happen if an idea were implemented. Outcomes included benefits, effects on users, behavioural changes, or measurable consequences. This was treated as elaboration because it extends the idea by specifying its implications (Kim, 2006; Chi, 2009).
- **Connection to another idea.** A connection to another idea was counted as an elaboration unit when a participant explicitly linked, combined, compared, or transferred content between ideas. It was not counted when ideas were merely mentioned side by side. This was grounded in collaborative problem-solving work showing that substantive group progress depends on participants taking up and building on one another's contributions (Barron, 2003).
- **Short agreement.** Short agreement included brief responses such as "yes", "I agree", or "that sounds good". These were excluded unless they contained additional task-relevant reasoning or development. This avoided treating social acknowledgement as substantive ideation content (Chi, 2009).
- **Repetition.** Repetition referred to repeating a previously stated idea, reason, or example without adding new task-relevant information. Repetition was excluded from both idea and elaboration counts to avoid double-counting the same contribution (Kim, 2006).
- **Filler.** Filler included hesitation markers, incomplete starts, conversational padding, or phrases used to hold the floor without adding task content. Filler was excluded because it reflected speech behaviour rather than contribution substantiveness.
- **Off-task remark.** Off-task remarks were comments unrelated to the ideation task, the proposed ideas, or the evaluation of those ideas. They were excluded from all contribution-substantiveness measures because the coding focused only on task-relevant idea content.
- **Unclear transcript fragment.** An unclear transcript fragment was excluded when the transcript did not provide enough information to determine whether the participant had introduced a new idea or elaboration. Ambiguous cases were coded conservatively to avoid over-interpreting the transcript.
- **Unsupported evaluation.** An unsupported evaluation was a judgement such as saying that an idea was good, bad, feasible, or unrealistic without explaining why. Unsupported evaluations were excluded from elaboration counts unless they introduced a criterion, reason, risk, constraint, or expected outcome. This follows the distinction between simple reaction and constructive contribution (Chi, 2009).