Training novices to think, or giving them LLMs? Evidence from an RCT

Hemanth Asirvatham
OpenAI

Chiara Betti
Bocconi University

Rachel Brown†
Independent Researcher

† Rachel Brown contributed to this work during her employment at OpenAI.

Arnaldo Camuffo
Bocconi University & ION Management Science Lab

Aaron Chatterji
OpenAI & Duke University

Chiara Fumagalli
Bocconi University

Alfonso Gambardella
Bocconi University & ION Management Science Lab

Myriam Mariani
Bocconi University

Abhinav Pandey
SDA-Bocconi & ION Management Science Lab

*Steve Ramos**
University of California, Berkeley

Giovanni Salvucci
Bocconi University

Vladimir Simic
SDA-Bocconi University

* Steve Ramos contributed to this work in his capacity as a paid contractor for OpenAI.

August 2026

Abstract

We study whether LLMs and cognitive skills influence the evaluation of novices' intellectual tasks, when these evaluations depend on standard and well-defined performance criteria. This is a common case (junior employees, students' exams) making our question relevant. In a 2×2 randomized control trial, we assign 1,053 first-year undergraduates in economics, management and finance to one of four conditions: a training on causal reasoning (as a relevant cognitive skill for intellectual tasks), ChatGPT Edu access (GPT-4o), both, or neither. We provide a real-world merchandising business problem and ask experts to evaluate the participants' solutions according to standard metrics in marketing. We find that causal reasoning changes how students think, leading to mechanism-based solutions and a greater diversity of ideas. However, causal reasoning does not improve evaluations, unlike ChatGPT. Greater textual coherence and a higher number of ideas explain about half of the ChatGPT effect, suggesting that, for novices facing well-defined problems, LLMs produce solutions that resemble expert recommendations not only in textual quality, but also in content. We also find that, when combined with causal training, ChatGPT increases the use of coherent logic, falsification logic, and the attempt to understand why (mechanisms). Overall, we find that human causal reasoning produces effects independently of LLMs, particularly on the diversity of ideas, and therefore it is worth cultivating causal reasoning in a world in which students use LLMs. Probably less appreciated is that the value of diverse ideas also requires an evaluation system that demands diversity rather than standard solutions.

1 Introduction

Do we still need to teach and learn standard cognitive skills in the age of artificial intelligence? If large language models (LLMs) can produce competent answers and solutions to problems, what should we still learn to retain a role in intellectual tasks? These are urgent questions that schools, universities, and organizations try to address while revising curricula, assessment criteria, and decisions to grant or restrict access to LLMs. Novices, such as young people, students, interns and entry-level workers, are the population for whom the answers to these questions matter the most. These people are in formative years in which AI may shape how and what they learn to approach intellectual problems. They have not yet acquired the domain knowledge that would help them evaluate what these new tools produce as an answer to specific problems, and, at the same time, are evaluated for what they produce, for instance by teachers on the learning outcomes, assignments and exams, or by managers for deciding hiring and career progressions (Garicano and Rayo, 2026).

In research, the phenomenon is recent, the paradigm is new and evidence on the effects of the new tools in combination with human cognitive skills is limited. Recent contributions document large heterogeneous effects of the use of AI in intellectual work, with complementarity or substitution effects depending on the type of task and the intellectual ability of the person performing it (e.g., Noy and Zhang, 2023; Dell'Acqua et al., 2026; Vaccaro et al., 2024; Bastani et al., 2025; Kestin et al., 2025). Typically, however, experimental studies manipulate the availability or type of an LLM tool and explore moderation effects by types of tasks and skills, with the latter being idiosyncratic to participants, coming from observational rather than experimental variation.

Our study focuses on an important category of novices, first-year university students who are new to business and economics subjects. It studies the effect of LLMs together with causal reasoning, a fundamental and teachable form of human learning and a natural candidate for the kind of capability that may remain human. Causal reasoning is the capacity to frame problems as structured chains of causes and effects, linking antecedents to outcomes through explicit mechanisms (Rubin, 1974; Angrist and Pischke, 2009; Pearl, 2009). We operationalize it as the ability to articulate causal explanations, specify why causal links operate, and identify falsifiable conditions under which they may not hold. Our design manipulates both LLM access and causal reasoning training, which allows us to examine the effect of each independently and, critically, in combination. The combination is what makes the teaching question answerable: it tells us not only whether a cognitive skill can be taught and whether it changes intellectual output, but whether it is still drawn upon when the option to outsource the task is available.

A preregistered 2×2 factorial randomized controlled trial conducted at a leading European university provides data from 1,053 first-year undergraduate students enrolled in 13 classes of the same first basic management course of three bachelor programs in economics & management (E&M), economics & finance (E&F), and management (M). We randomized at the classroom level within program to one of four conditions: causal reasoning training versus a placebo game, and ChatGPT Edu access versus none: Control, Causal, GPT, and Causal&GPT, with 249, 256, 197, and 351 participants respectively, oversampling the combined cell to gain power on the interaction.[1] We

[1] Overall, 90.5% of participants with GPT access declared compliance: 88% of those with GPT only, 92% of those with GPT and causal reasoning.

conducted the experiment during class time and did not inform students in advance, so that the participation in the experiment reflected regular class attendance. Participants then spent 45 minutes on a real-world business case, writing up to 180 words of consultant-style recommendations for increasing alumni awareness and usage of the university's merchandising shop, two established criteria in the marketing literature (Bendle et al., 2016). Performance was evaluated by twenty master's students serving as raters, three per solution, and separately against the recommendations of three domain experts. Figure 1 shows the experiment flow; Sections 1 through 4 of the Appendix provide the full design, rubrics, and measures. Section 5 in the Appendix provides balance checks.[2]

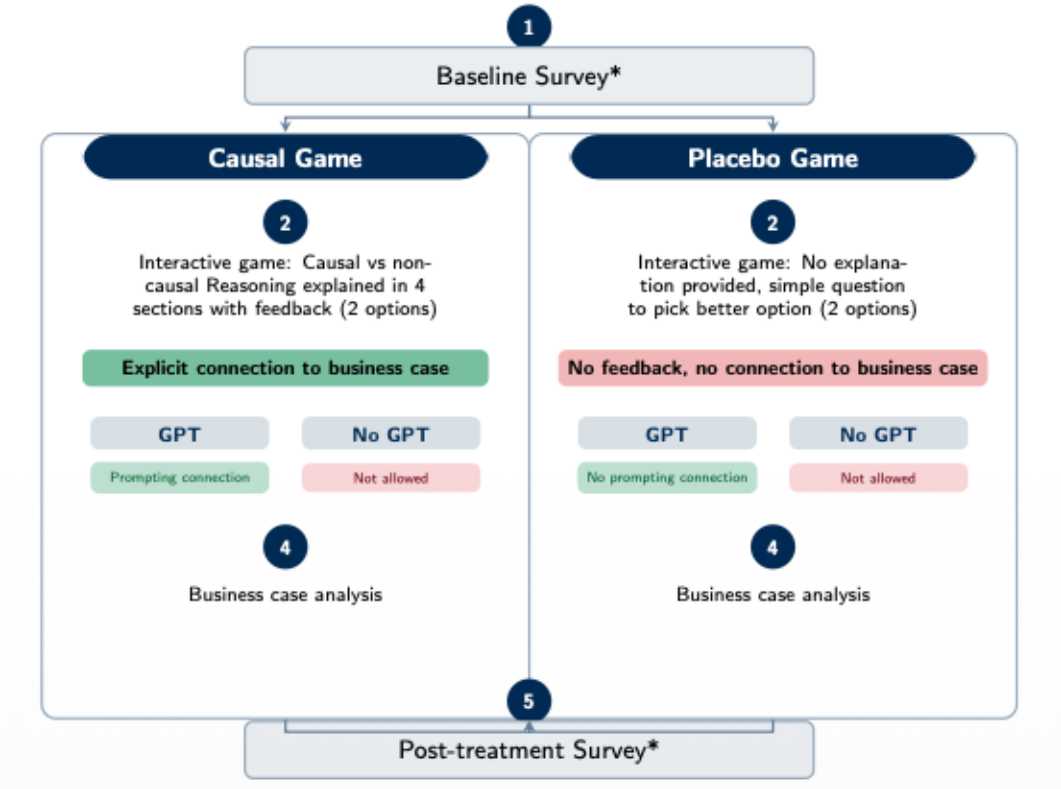

Figure 1: Experimental design

We find that the causal training changed how participants reason. Relative to control, it raises reasoning in terms of mechanism identification by about 0.5 standard deviations and use of falsification logic by about 0.8 standard deviations. It also increases the diversity of ideas: causal participants produce solutions containing more varied ideas, and ideas that are more distant from those of other participants. GPT changed the output in a different way. It improved the arguments' coherent logic and increased the number of ideas contained in each solution. On its own, however, it induced neither mechanism-based reasoning nor falsifiability, and it left the diversity of ideas largely unchanged.

Turning to evaluation performance, GPT is the only treatment that improves it. Access to ChatGPT

[2]To preserve authors' anonymity, we do not disclose the name of the university and the registration number of the pre-analysis plan with the American Economic Association's registry for randomized controlled trials.

raises the awareness and usage score, measured on a Likert scale 1-5, by an estimated 0.86 relative to the estimated score of 2.09 for the control group and makes the solutions closer to those provided by the domain experts. Thus, LLMs make novices resemble experts on standard tasks on which LLMs are well trained. Causal training alone has no effect, or a slightly negative one, and the same holds for the combined treatments. Post-double-selection lasso regressions show that number of ideas and coherent logic account for part of the GPT effect, but about half of it survives all controls, including those on textual features, indicating that GPT improves the substance of the solutions and not only their presentation. The same decomposition explains why causal reasoning does not pay: falsifiability and mechanism identification are negatively related to the evaluation score, and once they are included the coefficient on the causal treatment becomes positive. Evaluators reward a richer set of ideas within a solution but penalize solutions that depart from the typical solution space.

A final relevant result is the one on the number and diversity of ideas. The number of ideas composing a recommendation is higher for both causal-treated and GPT-assisted groups. The interaction term further benefits the number of ideas in the solution. For diversity, instead, causal training raises the diversity of ideas relative to the control sample. The availability of LLMs does not. However, it does not hamper the benefits from causal reasoning. The positive effect of causal training survives the availability of an LLM: causal training remains effective, even though the availability of LLMs does not add any advantage. Even though our design does not trace the cognitive process behind this result, we interpret this evidence as suggesting that participants exercise the trained skill without delegating the task to LLMs.

A direct implication of these results is that, if learned, a cognitive skill does not fail to be used even if a capable tool is within reach. A second implication concerns the way intellectual work is assessed. Human evaluation against codified criteria is pervasive: for example, professors grading exams, investors screening pitches, organizations assessing candidates for hiring, promotion, or project approval. When the task is well specified, the criteria are detailed in a standard rubric, and the model embodies the relevant knowledge, LLM access can help novices produce responses that score better than those produced by novices without AI, and better even than those of novices equipped with a relevant cognitive skill. This speaks about the need for cultivating diversity of thought, which may require training human (causal) reasoning and evaluation systems that demand diversity rather than standard solutions. The rubrics we used in the study penalizes distance from the conventional answer. If our desire is to go beyond conventional answers, then the binding constraint may lie in our capacity to demand and evaluate diverse, novel, and original ones, a long-standing difficulty that LLMs are making more visible and more urgent. Simply put, if we do not demand or value diversity in our assessments, the diversity produced by causal reasoning is not reflected in evaluation performance. LLMs have nothing to do with it: they do not enhance diversity compared to the group treated causally, but do not penalize it compared to the control group or in conjunction with the causal treatment.

Finally, the fact that novices trained in cognitive skills continue to apply them even when LLMs are available and, in doing so, produce more unconventional solutions, is encouraging for the development of the next generation of experts. If cognitively trained novices remain willing and able to tackle problems that LLMs have not encountered, were not trained on, or can learn to solve only as humans confront them, then closing today's novice–expert gap need not come at the

expense of developing tomorrow's experts.

2 Related research

Existing studies on the effects of humans collaborating with AI document substantial productivity gains in knowledge-intensive tasks. Noy and Zhang (2023) show that ChatGPT access among professionals reduces task completion time and raises output quality, with gains concentrated among lower-ability workers. This skill-equalizing pattern is replicated at scale by Brynjolfsson et al. (2025). Dell'Acqua et al. (2026) show that within the same knowledge workflow, AI assistance improved output quality on tasks inside its capability domain while reducing correct solutions by 19% on tasks outside it, a "jagged technological frontier". The meta-analytic baseline analysis by Vaccaro et al. (2024), reviewing 106 experimental studies, finds that human-AI combinations perform worse on average than the best standalone agent, with gains primarily in content-creation contexts and losses in tasks that imply making decisions. Notably, this literature addresses the human contribution through skill levels, domain expertise, cognitive style or other cognition characteristics that moderate returns to AI assistance (Krakowski et al., 2023; Bastani et al., 2025; Kestin et al., 2025).

The literature on AI in intellectual and creative work points to a task-specific form of complementarity (Chen and Chan, 2024). AI and humans bring distinct strengths to knowledge production, with human-AI solutions dominating human crowds on strategic viability (Boussioux et al., 2024). Also, while AI raises individual creative output, it narrows the collective idea space by anchoring users toward familiar solutions (Doshi and Hauser, 2024; Ju and Aral, 2026), and increases output homogeneity in co-written argumentative essays (Padmakumar and He, 2024). Evidence from over two million preprints across scientific disciplines shows that LLM adoption accelerates manuscript production and broadens literature discovery, but it reduces the probability of publishing (Kusumegi et al., 2025) or narrows science's focus (Hao et al., 2026). Agrawal et al. (2026) provide a theoretical account of the underlying complementarity: AI augments combinatorial search and hypothesis generation, while human judgment remains indispensable for abductive inference, contextual nuance, and causal reasoning in data-sparse environments.

On top of these contributions, we study learning and the performance of learning when using LLMs in a realistic professional problem. We do not simply infer the role of causal logic but provide a treatment to assess its implications in isolation and in combination with LLMs. Finally, we deal with a context – academic education – in which we need to better understand the implications of LLMs, particularly in relation to causal reasoning, one of the most relevant forms of learning.

3 Causal reasoning

Causal reasoning is a teachable form of human learning, used in training programs in schools and universities (Camuffo et al., 2020). We build on two complementary approaches to causal reasoning. Pearl (2009) formalizes causation as asymmetric determination in directed acyclic graphs (DAGs). Angrist and Pischke (2009) build on Rubin (1974)'s potential-outcomes framework and

operationalize causal claims through falsifiable null hypotheses: A causal statement is valid if it specifies the conditions under which it would not hold. Our study uses both frameworks and operationalizes causal reasoning through three constructs: (a) coherent logic, in the form of directed causal chains from antecedents to outcomes (A $\rightarrow$ B $\rightarrow$ C); (b) falsifiable statements, specified in terms of boundary conditions under which the proposed causal link might not hold; (c) mediating mechanisms explaining why the causal relationship operates.

In our experiment, the causal reasoning training took place through a structured game on all three constructs. The training was short and intense. We show participants real-world situations that require them to make decisions and choices according to a causal reasoning framework applied to the game. We divide the game into three sections in which we asked treated participants to answer questions about scenarios by making causal mechanisms explicit (explain why the causal relations work), using falsifiable statements, and employing a coherent logic narrative (describe the chain of causes and effects). We explicitly asked treated participants to use the tools learned during the training game. Participants not in the two causal conditions played an identical placebo game and answered the same questions but did not receive any instructions or reference to reasoning methods and they had no feedback on their answers.

We assessed the causal reasoning of the participant texts using a three-attribute rubric evaluating the use of *coherent logic, mechanisms*, and *falsifiability*. Section 3 of the Appendix describes the details of the rubric and the text evaluations.

4 Results

4.1 The causal treatment worked

Figure 2 (and Table A.9 in the Appendix) reports the estimated effects of the use of mechanisms, falsifiability and coherent logic with respect to the treatments. The graphs report the plot of the coefficients of the treatments divided by the standard deviation of the dependent variable. All the regressions in this and the following sections are OLS regressions with clustered standard errors and controls for the variables that fail the balance checks, as noted below each graph. Also, in all the regressions and figures the estimated coefficients of *Causal* and *GPT* are marginal effects with respect to the control group, while the estimated coefficient of the interaction *Causal X GPT* is the marginal effect with respect to the treatments alone. Therefore, a zero effect of the interaction means no effect on top of the effects of *Causal* or *GPT*, and a positive (or negative) effect is on top of these two effects.

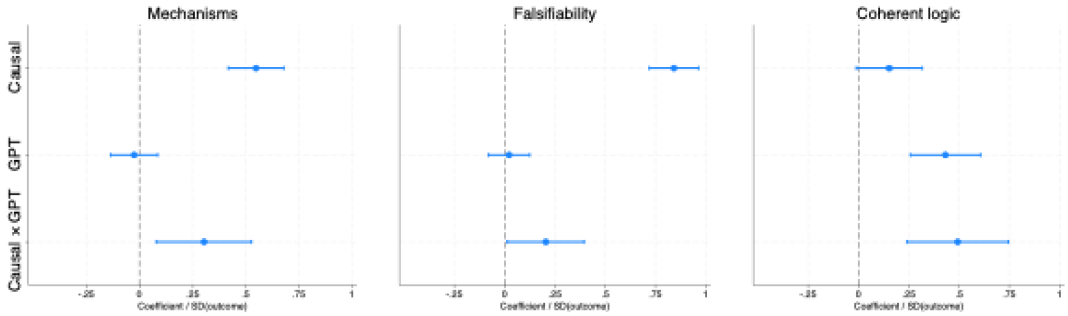

Figure 2: Estimated coefficients of Causal, GPT, and Causal X GPT on causal reasoning constructs

Note: N=1,053. OLS regression using clustered standard errors and controls for variables that failed the balance check. The figure reports regression coefficients normalized by the standard deviation of each dependent variable. The effect of the interaction is in addition to the effect of the treatments. The standard deviations of the dependent variables are, respectively, 0.437 (mechanisms), 1.027 (falsifiability) and 0.478 (coherent logic). Table A.9, columns (1), (2) and (3) in the Appendix report the estimated coefficients along with the variables used to cluster errors and controls.

The causal reasoning training was effective. Relative to the control group, it improves mechanism identification by about 0.55 standard deviations. GPT adds to mechanism identification only when combined with the causal condition. Falsifiability shows the largest effect: The causal condition increases the dependent variable by about 0.85 standard deviations. The interaction term is positive but small, around 0.20 standard deviations. We also show in Figure A.1 and Table A.9 (columns 4 and 5) of the Appendix that *Causal* participants took less time to complete the game and obtained a higher score than the control group.

Coherent logic, instead, increases mainly in the GPT condition. The effect is around 0.45 standard deviations. The interaction term is also positive, around 0.48 standard deviations. The effect of causal alone, though positive, is only 0.15 standard deviations. Thus, GPT is associated in particular with improvements in coherent logic. This improvement is reinforced when combined with the causal condition.

Overall, we find that the causal and GPT treatments are complementary on these three variables: GPT reinforces the positive effect of a causal treatment.

4.2 Performance: Awareness and Usage

After the causal treatment and placebo, participants worked on the same 45-minute business case administered through a Qualtrics survey interface. The task was realistic: participants were asked to write up to 180 words of a consultant-style recommendation for increasing alumni's *awareness* and *usage* of university merchandising. Awareness and usage are two widely established criteria in the marketing literature to promote products (Bendle et al., 2016). We gave participants in the ChatGPT conditions access to the tool through their institutional subscription in a separate browser tab. Since these are well-documented concepts in the marketing of products, LLMs are likely to be well-trained on this topic.

We evaluated individual performance by asking two types of human experts. For our main results, we use twenty students in the master's programs of the University. We randomly assigned each case to three of these raters and use the average of their evaluations. We picked master students drawn

primarily from management and marketing programs. As discussed in Section 2 of the Appendix, they have work experience, international academic experience through exchange programs, and outstanding academic curricula. They are highly qualified to evaluate a case about merchandising sales. We provided them with a clear definition of awareness and usage based on a detailed rubric from the marketing literature and we trained them by explaining the two terms and the rubric. Section 2 of the Appendix explains the performance measure, provides details about the rubric and the training of the evaluators and reports statistics about inter-rater reliability.

We then gathered expert evaluations from a marketing professor from the University, a manager of the University merchandising shop, and an expert University alumnus. We invited them, separately, for an interview, without revealing the objective of the meeting. During the interview we presented them with the same merchandising problem as those given to the experiment's participants and asked for their expert recommendation referring again to awareness and usage. To ensure that the solution came from their expertise rather than from the use of LLMs, we collected their feedback and wrote the text during the interview. We constructed an LLM-based performance measure of the distance between the texts of the participants and each one of the three experts. We take the texts of the experts as closer to ground truth based on awareness and usage.

Figure 3 plots the estimated coefficients of a regression that uses the experimental conditions as regressors and the 1-5 performance indicators (*awareness and usage*) averaging three raters and the two dimensions of awareness and usage, as discussed in Section 2 of the Appendix. The *GPT* treatment is the only driver of improvements in the *awareness and usage* evaluation score relative to the control group. In the no-controls specification, *GPT* increases the score by 0.834 points. With balance controls the estimate is 0.862 points, relative to the estimated score of 2.09 for the control group. Both are statistically significant at the 1% level. Thus, the *GPT* coefficient is substantively large. Table A.10 in the Appendix shows the tabulated results.

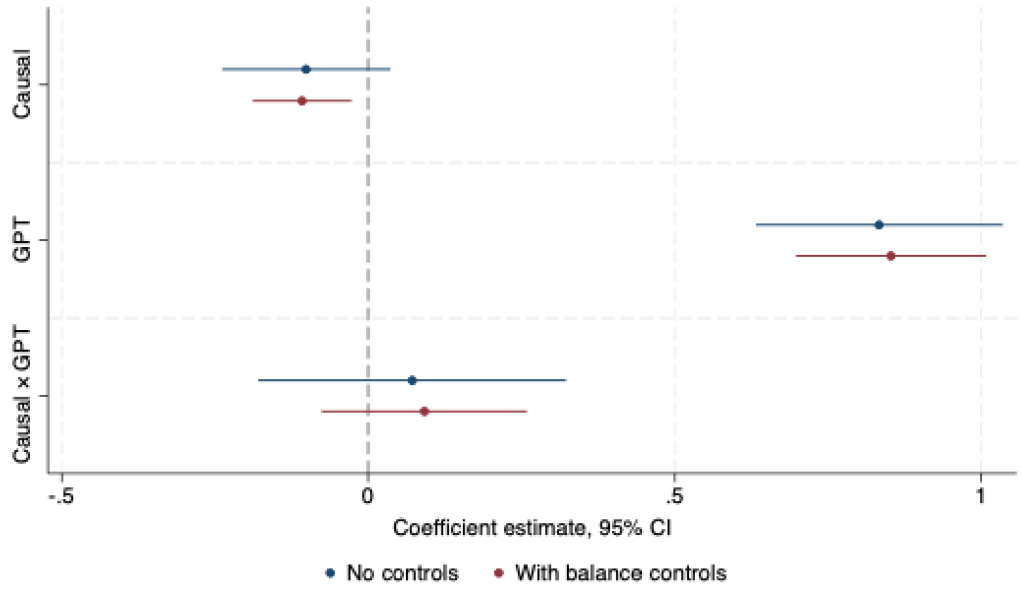

Figure 3: Estimated coefficients of Causal, GPT, and Causal X GPT on Awareness and Usage evaluation scores

Note: N=1,053. OLS regression using clustered standard errors and controls for variables that failed the balance check. The figure reports regression coefficients. The effect of the interaction is in addition to the effect of the treatments. The dependent variable mean is 2.6. Table A.10 in the Appendix reports the estimated coefficients along with the variables used to cluster errors and controls.

Causal alone has a slightly negative effect. There is no statistically meaningful evidence that combining *Causal* with *GPT* produces an additional effect, on top of using them alone**.**

In our experiment LLMs are well trained because marketing evaluations are a widely documented topic, while the novices are less familiar with it. In this context, ChatGPT closes the performance gap of non-experts facing evaluations based on well-specified criteria. Causal reasoning does not add much to the solution. As shown in the previous sections, it directs attention toward mechanisms or analytical structure that, as we will see in our lasso regressions below, are not rewarded because they are likely to reduce the alignment with what evaluators look for in this task.

Figure 4 plots the estimated coefficients of a set of three regressions that use as dependent variables three measures of similarity to the expert texts (*similarity*) derived from the distance indicators and the experimental conditions as regressors. Table A.11 in the Appendix shows the estimated coefficients. Again, in all three cases, the *GPT* treatment is the only significant regressor, reinforcing the evidence that GPT closes the gap between novices and experts.

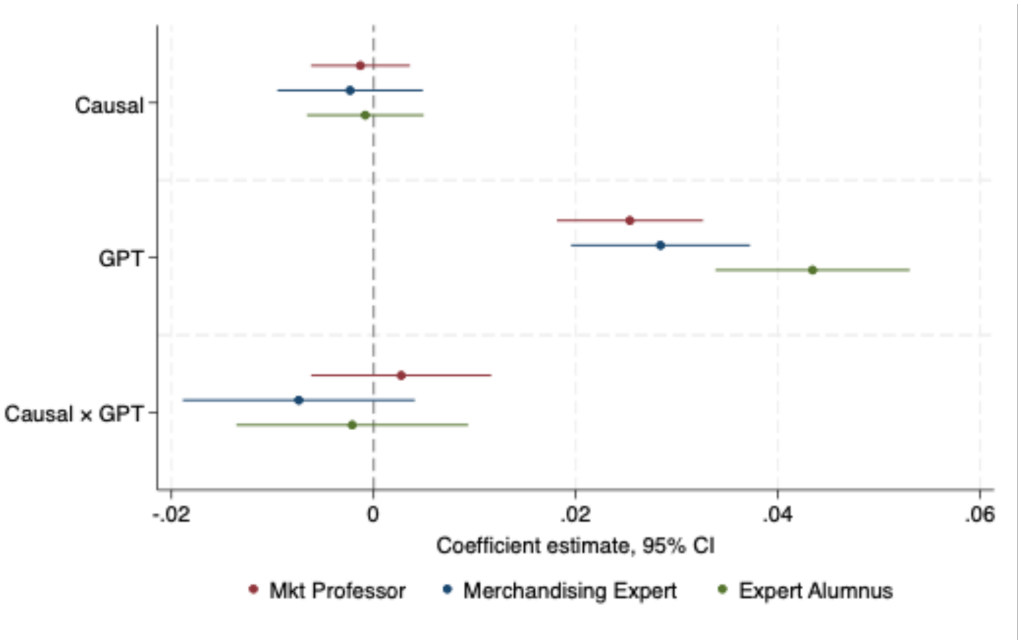

Figure 4: Estimated coefficients of Causal, GPT, and Causal X GPT on similarity between expert and participant solutions

Note: N=1,053. OLS regression using clustered standard errors and controls for a few variables that failed the balance check. The figure reports regression coefficients. The effect of the interaction is in addition to the effect of the treatments. The dependent variable mean is 0.79, 0.75 and 0.81 for similarity with Marketing Professor, Merchandising Expert and Expert Alumnus respectively. Table A.11 in the Appendix reports the estimated coefficients along with the variables used to cluster errors and as controls.

It is also worth noting that inter-rater reliability is higher for recommendations produced by the LLM-assisted group than for those produced by the control and causal-training groups. Because the novices' LLM-assisted recommendations more closely resemble those of the experts, raters appear to agree more when evaluating these texts. This evidence may reflect the content that LLMs prompt novices to include in their solutions. Therefore, we next examine the number and types of ideas contained in the recommendations and compare them across groups.

4.3 Ideas: Number and Diversity

We look at the ideas composing the recommendations produced by the students. Figure 5 reports standardized treatment effects on three dimensions: the number of ideas included in a solution; within-solution idea diversity, that is, for the same participant, the distance between the ideas that compose a solution; and between-solution idea diversity, that is, the distance between the core idea of a solution and the core ideas of the solutions provided by all the other participants. Section 4 in the Appendix describes the construction of these three variables. The coefficients in Figure 5 are expressed in standard deviation units of the corresponding outcome variable. Table A.12 in the Appendix shows the estimated coefficients.

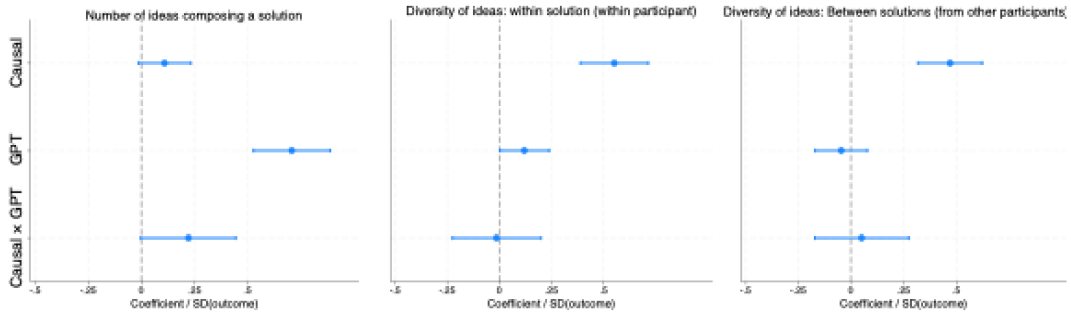

Figure 5: Estimated coefficients of Causal, GPT, and Causal X GPT on number and diversity of ideas

Note: N=1,053. OLS regression using clustered standard errors and controls for a few variables that failed the balance check. The figure reports regression coefficients normalized by the standard deviation of each dependent variable. The effect of the interaction is in addition to the effect of the treatments. Table A.12 in the Appendix reports the estimated coefficients along with the variables used to cluster errors and controls.

The results show that both *Causal* and *GPT* treatments are associated with changes in the structure of solutions, but their effects differ across outcomes. GPT has a large positive effect on the *number of ideas*, increasing the number of core ideas by more than one-half of a standard deviation. *Causal* also increases the number of ideas, although the effect is smaller. The same is true for the interaction between *Causal* and *GPT*, suggesting a weak complementarity between the two to increase the number of ideas beyond the separate effects.

For *within-solution diversity*, the *Causal* condition produces the strongest effect. Solutions generated under the causal condition contain more diverse ideas, with an effect of approximately one-half of a standard deviation of the dependent variable. *GPT* also has a positive but much smaller effect. The interaction between *Causal* and *GPT* is instead close to zero. Similarly, not only does causal reasoning increase the diversity of ideas pulled together by the individual participants in the same solution, but it is also the only treatment with a clear positive effect on *Between-solution diversity*. Solutions produced under the causal condition are more distant from solutions generated by other participants, indicating that causal reasoning encourages more distinctive or less conventional idea combinations. By contrast, *GPT* alone and the interaction between *Causal* and *GPT* are not distinguishable from zero. Taken together, these results suggest that GPT expands the volume of ideas that people combine to produce a solution. However, it is causal reasoning that increases the diversity of the ideas, both within the same solution and across participant ideas. Importantly, the availability of GPT is not detrimental to the generation of diverse ideas. This is true also for people trained with causal reasoning when they have the option to use GPT.

4.4 Bringing the evidence together

To pull all our findings together, we performed a set of post-double-selection (PDS) Lasso regressions that always retain the three treatment variables, while the algorithm selects the controls (among those listed below Table A.13 and Table A.14 of the Appendix). Figure 6 plots the estimated results from these regressions and, by comparison, those from the baseline OLS regression (with balance controls) whose results are shown also in Figure 3. *Awareness and usage* evaluation

score is the dependent variable. Table A.13 in the Appendix shows the tabulated results.

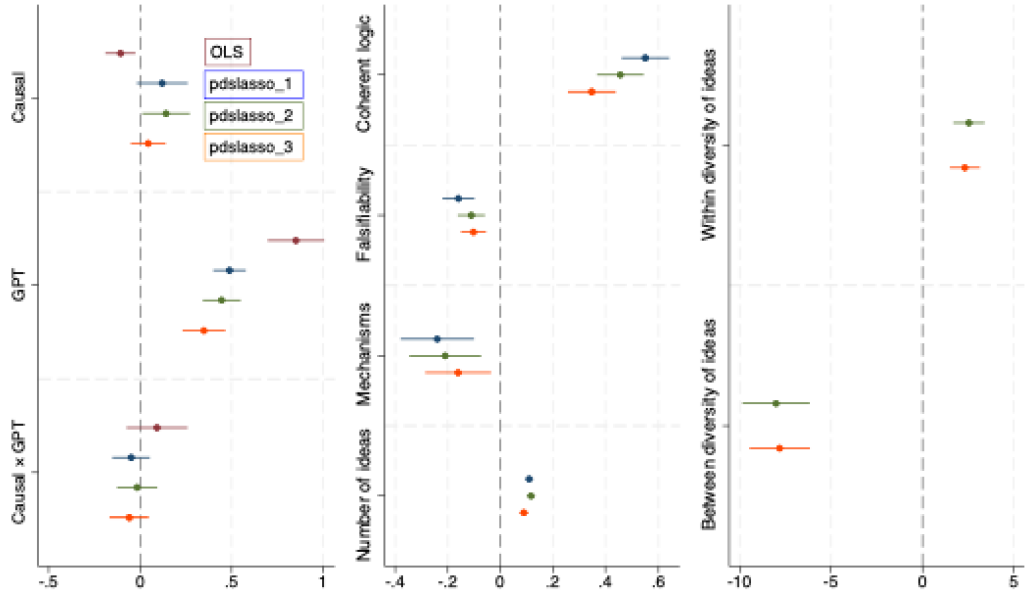

Figure 6: Estimated coefficients on Awareness and Usage Evaluation scores: OLS and PDS Lasso regressions

Note: N=1,053. OLS regression using clustered standard errors and controls for a few variables that failed the balance check. The figure reports regression coefficients. The effect of the interaction is in addition to the effect of the treatments. The dependent variable mean is 2.6. Table A.13 in the Appendix reports the estimated coefficients along with the variables used to cluster errors and controls.

The three PDS Lasso models add controls progressively. PDS Lasso 1 controls for content features: coherent logic, falsifiability, mechanisms, and number of core ideas. PDS Lasso 2 adds diversity/distance measures: within-solution diversity of ideas and between-solution diversity of ideas. PDS Lasso 3 further expands the control set to include writing style characteristics such as sentence count, average sentence length, and syntactic features (described in Section 4 of the Appendix).

The comparison across models shows whether the estimated treatment effects survive after accounting for the observable characteristics of the submitted solutions.

The results show that GPT has a large, positive, and statistically significant effect on *awareness and usage* across all PDS Lasso models. The GPT coefficient is the largest in the OLS regression 0.862, and drops to 0.490 in PDS Lasso 1, 0.448 in PDS Lasso 2, and 0.362 in PDS Lasso 3. The coefficient declines as we add more controls, but it remains positive and statistically significant at the 1% level throughout. This indicates that GPT is effective on the *awareness and usage* task, and that part of its advantage operates through producing more coherent, richer, and more polished responses. The GPT coefficient declines when we include the causal reasoning variables, with coherent logic being highly correlated with GPT and when we control for the writing style variables (from PDS Lasso 2 to PDS Lasso 3). It is worth noting that about half of the GPT effect remains after controlling for text features, number and diversity of ideas, causal reasoning

characteristics, etc., suggesting that GPT itself increases the quality of the solutions produced.

The negative coefficient of causal reasoning estimated from the OLS regression becomes positive and loses statistical significance when we add text controls in PDS Lasso 3. Similarly, the *Causal-GPT* interaction is consistently small and statistically insignificant below standard levels across the three PDS Lasso specifications, and smaller than in the OLS regression.

The regressors added stepwise explain why. *Coherent logic* is positive and highly significant in all models, though its coefficient declines from 0.547 to 0.355 as we introduce additional controls. The *Number of ideas* is also positive and highly significant throughout, with coefficients around 0.092-0.116. As shown earlier, both variables are correlated with the use of GPT and explain part of the GPT effect in the baseline regressions.

Differently, several variables associated with causal reasoning are negatively related to performance. *Falsifiability* is negative and highly significant in all models, with coefficients between –0.159 and –0.109. *Mechanisms* are also negative and significant, with coefficients ranging from –0.242 to –0.166. This suggests that evaluators did not reward answers that emphasized testability or causal mechanisms and may have penalized them relative to more conventional *awareness and usage* responses. The negative effect of *Causal* vanishes with the inclusion of these variables, suggesting that they capture the negative effect of *Causal*.

The diversity variables tell a similar story. *Within-solution diversity of ideas* is positive, indicating that evaluators reward answers that contain a richer set of ideas within the solution. However, *Between-solution diversity of ideas* is negative, implying that answers that are more distant from the typical solution space receive lower scores. In other words, novelty or distance from the standard answer is not rewarded with the current rubric, even if it may reflect more original solutions.

Finally, the text variables added in PDS Lasso 3 reduce both the coefficient of *GPT* and *Causal*, suggesting that part of the effects of both goes through the specificities of the text of the recommendations, on top of other substantive factors.

Table A.14 in the Appendix re-estimates the PDS regressions for the similarity to expert measures as dependent variables. The results largely confirm those shown in Figure 6 and Table A.13. In all models GPT retains a positive and statistically significant coefficient. However, the size of the coefficient is smaller than in the OLS regression in Figure 4 and Table A.11, falling to about a third of the OLS coefficient in the third specification of each set, when we add controls for text characteristics. Again, these results confirm that GPT is effective on the *awareness and usage* task, making novices more similar to expert evaluators, and that part of the advantage operates through producing more coherent, richer, and more polished responses. A third of the effect of GPT, though, remains even after controlling for all these factors that GPT may operate through, confirming that it increases the quality of the solutions produced, above and beyond a more coherent and polished text.

The negative coefficient of causal reasoning estimated from the OLS regression becomes positive and remains statistically not significant below standard levels in most PDS Lasso regressions, whereas the *Causal-GPT* interaction is consistently negative and gains statistical significance in some of the specifications. Again, as for the master students' evaluations, *Coherent logic* and *Number of ideas* are positive and statistically significant throughout, while variables associated with causal reasoning are negatively related to performance. Finally, *Within-solution diversity of*

ideas is positive, while *Between-solution diversity of ideas* is negative, confirming the penalty associated with novelty with the current evaluation rubric.

5 Conclusions

Our experiment shows that LLMs close the novice-expert gap and perform well on problems with well-defined solution metrics. Access to ChatGPT raises the evaluation scores of participants' recommendations and moves their solutions closer to those produced by domain experts. Causal reasoning training did neither.

This type of assessment that evaluates codified output against detailed criteria is frequent in the real world, for example professors grading exams, investors screening pitches, and organizations deciding on hiring, promotion, or project approval. In these cases, our study indicates that the advantage of LLMs is not only cosmetic: part of it survives all textual controls and coherent logic features, suggesting that LLMs improve the substance of what novices produce and not only the way the content is presented. In addition, by bringing the recommendations of less expert evaluators closer to those of domain experts, LLM-assisted texts are easier to evaluate, creating less disagreement among raters.

The picture reverses when we turn from performance to the diversity of ideas. Causal reasoning training is effective in raising mechanism identification and falsifiability, and it is the only treatment that produces more diverse ideas, both within a participant's solution and relative to the solutions of other participants. Yet these are precisely the attributes that the evaluation rubric penalized: mechanisms and falsifiability are negatively associated with the evaluations; similarly, evaluators mark down solutions that are further from the typical solution space. Distance from the standard, in other words, is not rewarded by the rubric, even where it may reflect a more original answer. This explains the null effect of the causal treatment on measured performance without implying that the training failed. The assessment system simply does not value its effect on the solutions and evaluators are in higher disagreement when rating of these solutions than for the LLM-assisted ones.

Interestingly, the combined effect of GPT and causal training is close to zero on the diversity of ideas indicators. Thus, relative to *Causal* alone, adding the availability of LLMs does not increase diversity, but it does not kill it either. We interpret this result as suggesting that the possibility to exercise a trained skill is not crowded out by the option to use LLMs. This finding suggests that it matters to teach people to think even in the age of AI, as relevant cognitive skills do not go unused when a capable tool is within reach. This is reinforced by our finding that causal training and GPT have complementary effects on three key categories of causal thinking: coherent logic (that is, thinking in terms of a clear causal chain A B C), falsifiability (that is, thinking in terms of counterfactuals), and mechanism identification (that is, understanding why). In this respect, training and tools are becoming increasingly inseparable decisions, and evidence that a skill can be taught, and its way of thinking is complementary to the tools, is also evidence that it will be drawn upon for tasks that the tool is not trained on.

Three implications follow. First, where the task is well specified and the model embodies the rele-

vant knowledge, well-trained LLMs outperform novices working alone and even novices equipped with a relevant cognitive skill. Second, if we want diversity of solutions rather than competent standard ones, we may want to cultivate human (causal) reasoning, irrespective of whether LLMs are available. Third, diversity requires an evaluation system that demands diversity rather than conformity. The binding constraint may lie less in our capacity to generate answers than in our capacity to demand and evaluate diverse, novel, and original ones. This is a long-standing issue that LLMs are making more visible and more urgent. Such a demand would also motivate humans to learn cognitive skills (such as causal reasoning) and to train LLMs accordingly.

These results leave open a question our design cannot settle. We observe the performance of learning, and we can decompose it: About half of the LLM advantage operates through clearer, richer, better organized text, and about half through the substance of the recommendations. Whether the second component reflects knowledge that participants acquired and retained, or output they procured without acquiring anything, we cannot say.

References

Agrawal, A.K., J. McHale, and A. Oettl (2026). *AI in Science. NBER Working Paper 34953, www.nber.org.*

Angrist, J.D. and J.S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton NJ.*

Bastani, H., O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı, and R. Mariman (2025). *Generative AI without Guardrails Can Harm Learning: Evidence from High School Mathematics. Proceedings of the National Academy of Sciences (PNAS) 122(26).*

Bendle, N.T., P.W. Farris, P.E. Pfeifer, and D.J. Reibstein (2016). *Marketing Metrics: The Manager's Guide to Measuring Marketing Performance. Pearson, Upper Saddle River NJ.*

Boussioux, L., J.N. Lane, M. Zhang, V. Jacimovic, and K.R. Lakhani (2024). *The Crowdless Future? Generative AI and Creative Problem-Solving. Organization Science 35(5): 1589-1607.*

Brynjolfsson, E., D. Li, and L. Raymond (2025). *Generative AI at Work. The Quarterly Journal of Economics 140(2): 889–942.*

Camuffo, A., A. Cordova, A. Gambardella, and C. Spina (2020). *A Scientific Approach to Entrepreneurial Decision Making: Evidence from a Randomized Control Trial. Management Science 66(2): 564-586.*

Camuffo, A., A. Gambardella, S. Kazemi, and A. Pandey (2026). *Beyond Black Boxes: Designing and Testing Agentic AI Systems for Strategy. Strategy Science 11(1): 137-156.*

Chen, Z. and J. Chan (2024). *Large Language Model in Creative Work: The Role of Collaboration Modality and User Expertise. Management Science 70(12): 9101-9117.*

Dell'Acqua, F., E. McFowland III, E. Mollick, H. Lifshitz, K.C. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K.R. Lakhani (2026). *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. Organization Science 37(2): 403-423.*

Doshi, A.R. and O.P. Hauser (2024). *Generative AI Enhances Individual Creativity But Reduces the Collective Diversity of Novel Content. Science Advances 10(28).*

Garicano, L. and L. Rayo (2026). *Training in the Age of AI: A Theory of Career Viability. CEPR Discussion Paper.*

Hao, Q., F. Xu, Y. Li, and J. Evans (2026). *Artificial Intelligence Tools Expand Scientists' Impact But Contract Science's Focus. Nature 649(8099): 1237-1243.*

Ju, H. and S. Aral (2026). *Collaborating with AI Agents: Field Experiments on Teamwork, Productivity, and Performance. arXiv, https://arxiv.org/abs/2503.18238.*

Kestin, G., K. Miller, A. Klales, T. Milbourne, and G. Ponti (2025). *AI Tutoring Outperforms In-Class Active Learning: An RCT Introducing a Novel Research-Based Design in an Authentic Educational Setting. Nature: Scientific Reports 15, 17458.*

Krakowski, S., J. Luger, and S. Raisch (2023). *Artificial Intelligence and the Changing Sources of Competitive Advantage. Strategic Management Journal 44(6): 1425–1452.*

Kusumegi, K., X. Yang, P. Ginsparg, M. de Vaan, T. Stuart, and Y. Yin (2025). *Scientific Production in the Era of Large Language Models. Science 390: 1240-1243.*

Noy, S. and W. Zhang (2023). *Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. Science 381: 187-192.*

Padmakumar, V. and H. He (2024). *Does Writing with Language Models Reduce Content Diversity? International Conference on Learning Representations, 642-669. https://arxiv.org/abs/2309.05196.*

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge University Press, Cambridge UK.*

Rubin, D.B. (1974). *Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. Journal of Educational Psychology 66(5): 688–701.*

Vaccaro, M., A. Almaatouq, and T. Malone (2024). *When Combinations of Humans and AI are Useful: A Systematic Review and Meta-Analysis. Nature Human Behaviour 8: 2293–2303.*

Appendix

This document contains the following additional material:

- *Section 1: Experimental design and flow*
- *Section 2: Performance evaluation: rubric, evaluators' training and inter-rater reliability*
- *Section 3: Causal reasoning measurement*
- *Section 4: Variables construction*
- *Section 5: Descriptive statistics and balance table*
- *Section 6: Tables with estimated results*

Section 1: Experimental design and flow

The study was pre-registered with the AEA RCT Registry [*link currently not available to preserve anonymity*] prior to data collection and received ethics approval from the University's Ethics Committee.

We conducted a 2×2 between-subjects experiment in November 2025 with first-year undergraduate students of a leading European University in the social sciences. The two treatments were (i) a causal reasoning training intervention (treatment vs. placebo) and (ii) access to ChatGPT (allowed vs. not allowed) for a recommendation to a business challenge. The sample comprised 1,053 students from 13 class sections (hereafter, classes) of the same introductory Management course, spanning three bachelor programs: 4 classes from a bachelor program in economics and management (E&M), 4 classes from a bachelor program in economics and finance (E&F), and 5 classes of a bachelor program in management (M). E&M and E&F are taught in English, M in the national language of the University. Within each program, students are assigned to classes by the university administration upon enrolment through an administrative process that is independent of student choice or past academic performance and does not anticipate our experiment. The treatments of our research design are randomized at the class level: we assign one intact class per program to each of the four arms. Only in the case of M, for which we have 5 classes, two classes (instead of one) were assigned to the joint treatment arm to increase statistical power in the condition of primary theoretical interest. Materials were administered in English for E&M and E&F and in the national language for M. In all the 13 classes the experiment took place within two consecutive days.

Each session lasted for approximately 45 minutes and was administered during regular class hours under proctor supervision. Completion times were logged, with students allowed to finish earlier than the total allocated time, or to stay a few minutes longer if needed to complete the assignment.

The flow of the session is as follows. After providing consent, participants completed a baseline survey covering high school track, a thinking-style scenario, and a 7-point Likert battery of

questions measuring attitudes toward AI and self-assessed knowledge about business strategy (see Section 4 of this Appendix for the list of variables). After completing the survey, participants proceeded to a "business game" that provided the training intervention for the selected treated group and a placebo game for the control group. The training took the form of a progressive pedagogical game comprising 12 binary-choice questions organized into four sections explaining and practicing the following concepts: causal thinking, coherent logic, falsifiability and mechanism identification. Each section introduced the concept's underlying principle through worked examples before presenting practice questions. Every response was followed by feedback explaining the correct answer, regardless of whether the answer provided by the student was correct or wrong. A concluding summary illustrated how the four principles combine in a single example proposal. The term "causal reasoning" was never used; the game described its content as "critical thinking skills" and to avoid anchoring participants on business ideas, questions were framed around policy interventions in a hypothetical city. In the placebo game, instead, participants simply selected answers from a list of options related to the same 12 questions based on the same hypothetical city case, but they were not presented with any scaffolding, example, or feedback to their answers. So, they were exposed to the same case, but with no training of explanation of the causal reasoning concepts.

After the causal reasoning treatment/placebo game phase, all participants received the same business case related to the University's merchandising division. They were given operational data on sales channels, pricing, and the 144,000+ alumni base and asked to provide a recommendation to increase Alumni's awareness and usage with the University branded merchandise. The task was completed in the form of a written recommendation of approximately 180 words. The participants were prompted very clearly in the text of the case that their advice had to aim at increasing awareness and usage of the University merchandise with the Alumni.

The participants in the AI-access arms could use ChatGPT Edu during the task (GPT-4o), while those in the non-AI arms did not. Participants in the causal reasoning training arm additionally received the prompt "Remember the skills you learned during the game" before beginning the task. The session closed with a post-treatment survey that repeated the baseline attitude items, recorded self-reported strategic approach and ChatGPT usage, and invited participants in the AI arms to share their conversation logs.

The recommendations were then assessed by a set of evaluators on the notion of awareness and usage from the marketing literature and were given a detailed rubric with the criteria based on these concepts (see Section 2 of this Appendix).

Section 2: Performance evaluation: rubric, evaluators' training and inter-rater reliability

2.1. Definition and Measurement

Our measure of performance is the effectiveness of recommended strategies to move Alumni from initial merchandise awareness to active usage. The construct is adapted from the AAU (Awareness,

Attitudes, Usage) marketing framework (Bendle et al., 2016), with attitudes deliberately omitted to isolate the two behavioural endpoints most directly relevant to the merchandising outcomes desired by the merchandising division at Bocconi.

The unit of analysis is the complete strategy recommendation text developed by each participant. Each text received a score for each of the two attributes, following Boyatzis (1998)'s framework for qualitative content analysis, where the object represents the bounded corpus being examined while units represent the specific elements being coded. Holistic assessment of the text prevents fragmentation of the awareness-to-usage funnel logic that coherent merchandising strategies must articulate. The rubric employs qualitative content analysis focused exclusively on manifest content. A structured framework transforms qualitative observations into quantifiable scores via a 5-point (1-5) Likert scale with operationalized anchoring criteria for the two distinct attributes of awareness and usage. Each pre-registered scale level is operationalized through explicit anchored descriptors and worked examples, with the rater training in §2.3 ensuring consistent application. Scale levels within each attribute are orthogonal and mutually exclusive, with each level representing a progressively higher degree of performance on that dimension.

2.2. Attribute Structure

Attribute 1: Awareness

Definition: The effectiveness of strategies designed to ensure that alumni notice, recognize, and recall the existence of the merchandise program and its products. Key indicators are:

- Communication reach: breadth and depth of channel deployment (email, social media, event activation, direct mail). Alumni-specific channels (LinkedIn, reunion events, alumni newsletters, class WhatsApp groups, professional networks) are given higher weight than generic channels.
- Message distinctiveness: concrete tactical elements that enhance recall rather than abstract memorability. Examples include graduation-anniversary campaigns, limited-edition decade collections, professional product positioning, specific timing strategies, and unique value propositions.

Table A.1: Awareness scoring scale.

Level	Description	Example
5 — Complete Awareness	Complete breadth and depth of channel deployment with strong emphasis on alumni-specific channels. Complete distinctiveness through multiple concrete tactical elements. Includes scale-amplification strategies (referral programs, alumni ambassadors, viral mechanisms). Alumni attention and recall fully assured.	*Deploy multi-channel campaign across LinkedIn alumni groups, reunion events, and 25th/50th anniversary outreach. Feature limited-edition decade collections with professional positioning. Implement alumni ambassador program with referral incentives to amplify reach across global network.*
4 — Substantial Awareness	Substantial breadth and depth of channel deployment with a good mix of alumni-specific and general channels. Substantial distinctiveness through several concrete tactical elements. Alumni attention and recall largely assured.	*Launch campaign through alumni newsletter, Facebook groups, and homecoming events. Create graduation-year-specific collections with professional designs. Use targeted email sequences based on graduation decade.*
3 — Moderate Awareness	Moderate breadth and depth of channel deployment with some channel diversity. Moderate distinctiveness through some concrete tactical elements. Alumni attention and recall somewhat assured.	*Promote through email and social media. Feature special alumni designs. Use periodic newsletters to maintain visibility.*
2 — Limited Awareness	Limited breadth and depth of channel deployment with minimal channels. Limited distinctiveness through few concrete tactical elements. Alumni attention and recall minimally assured.	*Send email announcements about new merchandise. Post on social media occasionally.*
1 — Absent Awareness	No breadth or depth of channel deployment. No concrete tactical elements to drive recall. Alumni attention and recall not assured.	*Make merchandise available online.*

Attribute 2: Usage

Definition: The effectiveness of strategies that drive actual purchasing behaviour, facilitate transaction completion, and encourage ongoing engagement with institutional merchandise. Key indicators are:

- Friction reduction: Tactics that simplify purchasing for the alumni context (digital/mobile optimization, corporate billing integration, bulk ordering for gifts/teams, pre-orders at reunion events, subscription programs, gift shipping services, personalization tools, interna-

tional shipping solutions).

- Transaction incentives: Promotional offers, discounts, bundles, or limited-time opportunities that motivate immediate purchase.
- Behavioural triggers: Professional identity reinforcement, milestone connections (anniversaries, promotions), exclusivity/scarcity, social proof from successful peers, nostalgia paired with contemporary relevance.
- Engagement mechanisms: Strategies encouraging repeat purchase, merchandise display, or ongoing program participation.

Table A.2: Usage scoring scale.

Level	Description	Example
5 — Complete Usage	Complete friction reduction with multiple tactics addressing alumni-specific contexts. Strong behavioral triggers connecting to professional identity and alumni life. Clear acknowledgment and solutions for alumni barriers (geography, relevance, professional use). Robust engagement mechanisms.	*Offer mobile checkout with corporate billing integration and international shipping. Launch professional desk accessory line timed to promotion season with 25th anniversary edition. Provide bulk ordering for corporate gifts with personalization. Implement quarterly subscription box featuring limited items. Address geographic barriers with global fulfilment centers.*
4 — Substantial Usage	Substantial friction reduction with several relevant tactics. Good behavioral triggers addressing alumni needs. Some acknowledgment of alumni-specific barriers. Clear engagement approach.	*Streamline checkout with saved profiles and gift shipping. Create professional merchandise line for office display. Offer reunion-timed discounts and bundle deals. Implement loyalty program with milestone rewards.*
3 — Moderate Usage	Moderate friction reduction with some specific tactics. Basic behavioral triggers present. Some consideration of alumni context and needs.	*Provide easy online ordering and gift shipping options. Feature alumni-exclusive items with graduation-year designs. Send anniversary-timed promotions.*
2 — Limited Usage	Limited friction reduction with unclear or minimal tactics. Few behavioral triggers. Limited consideration of alumni-specific needs.	*Make online purchasing available. Offer some discounts occasionally.*
1 — Absent Usage	No friction reduction; no tactics employed. No consideration of alumni-specific contexts or barriers.	*Merchandise can be purchased through the campus bookstore.*

Overall Score

The overall Alumni Merchandising Performance score is the unweighted average of the two attribute scores provided by each evaluator: Performance = (Awareness + Usage) / 2.

2.3. Evaluators' Recruitment

The evaluators' pool is composed of 20 students recruited through an open call distributed via departmental mailing lists. They were master's students at leading European University, predominantly first-year and enrolled in management-oriented programmes: 60% in International Management, 20% in Arts, Culture, Media and Entertainment management, 10% in Marketing Management, and 10% in other Economics and Management tracks. Academically the group was strong: 45% had graduated from their bachelor's degree cum laude or with first-class honours, and 75% had completed at least one period of international study, whether an exchange semester or a full prior degree earned abroad. Recruitment materials described the task in general terms, namely, as a structured rating exercise on business consulting. All evaluators were blind to the experimental conditions, the assignment of stimuli to conditions, and the study hypothesis. No information regarding the predicted direction of effects was conveyed at any point. Prospective evaluators were screened for prior involvement in related projects, and any individuals with prior exposure to the stimulus set or the research questions were excluded from participation.

Training was delivered in two stages. The first stage consisted of a two-hour initial session in which evaluators were introduced to the rating rubric, the operational definitions of each construct, and the scoring scale. The session included a guided walkthrough of annotated example stimuli that were not drawn from the experimental set, followed by a supervised practice block in which evaluators applied the rubric independently and then discussed their ratings as a group under the guidance of the training facilitator. The second stage, conducted on a subsequent day, consisted of a one-hour recalibration session in which evaluators reviewed a fresh batch of calibration examples, compared their ratings against a reference key, and discussed sources of disagreement. The recalibration session was intended to reduce drift, surface systematic biases, and align interpretation of the rubric's edge cases before the start of the main rating phase.

Throughout both the training and rating phases, evaluators had continuous access to the rating rubric, a written codebook elaborating each construct with examples, and a full video recording of the initial practice session for reference. These materials were made available to support consistent application of the rubric and to allow evaluators to resolve ambiguities by consulting the codebook rather than relying on memory or peer consultation during active rating.

All training, calibration, and rating sessions were conducted on-site at the University under standardised conditions. Evaluators were not permitted to take stimulus materials off-site, nor to discuss specific stimuli or their ratings with one another outside of the structured calibration discussions. Stimuli were presented in a randomised order that varied across evaluators, and any identifying metadata that could have signalled condition membership was removed prior to presentation. The training facilitator was the only individual present during training who had knowledge of the experimental design, and the facilitator's role was limited to procedural guidance.

The total workload per evaluator was approximately ten hours, inclusive of the training and re-calibration sessions. Payment was not contingent on rating outcomes or on agreement among evaluators.

In addition to the evaluations provided by the twenty master students, we asked for more senior expert support. We called a marketing professor at the University, a manager of the merchandising shop, and a University alumnus expert on the topic, for an interview without anticipating the content. During the interview we presented each of the three experts (separately) to provide their own recommendation to the business case, to increase awareness and usage. We asked them to read the case and to write their text during the interview. The three expert recommendations averaged approximately 208 words in English (range 196–217; approximately 222 words in the manually translated Italian versions). In this way, we ensured that they did not use LLMs. We then compared their recommendations with those provided by the participants in the experiment, and computed an indicator of performance measured in terms of distance between experts' recommendations and students' recommendations.

2.4. Performance Evaluations

Each of the 1,053 case responses received three independent assessments by three of the twenty trained evaluators, assigned in a balanced design (approximately 158–159 responses per evaluator). We distributed cases such that each evaluator rated the same case only once, and we assigned evaluators preserving balance across the four experimental arms within each evaluator's individual workload, mirroring the balance in the full case set of 1,053. We retained complete sampling logs, including case identifiers, evaluator identifiers, assignment timestamps, and arm membership, for all 3,159 ratings (1,053 X 3).

Each evaluator scored the two dimensions of awareness and usage on the 1-5 Likert scale. We first compute the evaluator's general score as the (within evaluator) average between awareness and usage scores. Then, the total score of each recommendation is computed as the average of the general score across the three evaluators. Table A.3 shows inter-reliability statistics.

Table A.3: Evaluators' inter-rater reliability

	Overall	*Control*	*Causal*	*GPT*	*Causal X GPT*	N
Interclass correlation (20 raters, 3 per response)	0.544	0.275	0.322	0.534	0.493	1,053
Asymptotic s.e.	0.017	0.041	0.040	0.040	0.031	
Convergent validity: r = 0.808 (Overall)						

Based on the comparison between senior experts' recommendations and students' recommendations, we derived 3 measures of performance. Each expert distance measures the semantic dissimilarity between each student's response and three expert-written model solutions, computed using cosine distance in embedding space.

The technical specifications for computing this distance measures are as follows:

- Embedding model: text-embedding-3-large (OpenAI)

- Language matching: Each student is compared to the expert solution in their preferred survey language (English or Italian). Expert solutions were manually translated to Italian to avoid cross-lingual embedding confounds.
- Expert profiles. Three expert solutions were authored for the merchandising case: (1) Marketing professor; (2) Merchandising manager; (3) University Alumnus.
- Pre-processing: no stop-word removal or pre-processing applied — text-embedding-3-large produces contextualised embeddings where common words have minimal influence on cosine similarity.

Based on the comparison between the expert texts and the respondent texts, we constructed three distance measures:

- *Distance from Mkt Professor*: Cosine distance (0–1) between student embedding and Professor solution embedding. The higher the value, the more distance the solution of the student from the Professor.
- *Distance from merchandising manager*: Cosine distance (0–1) between student embedding and Merchandising office solution embedding. The higher the value, the more distance the solution of the student from the Merchandising office expert.
- *Distance from alumnus*: Cosine distance (0–1) between student embedding and University Alumnus solution embedding. The higher the value, the more distance the solution of the student from the Alumnus.
- We use 1 – distance (so that higher values imply more similar solutions) as dependent variables in our analysis and label them as *Similarity of participant solution with Mkt Professor solution, Similarity of participant solution with Merchandising expert, Similarity of participant solution with Expert Alumnus.*

Section 3: Causal reasoning measurement

3.1. Measurement

We measured causal reasoning by using LLM to score the presence of three attributes in the business case text: coherent logic, mechanism identification, and falsifiability. These attributes combine the formalization of causation of Pearl (2009) (i.e., asymmetric determination in directed acyclic graphs) and Angrist and Pischke (2009) (i.e., causal claims through falsifiable null hypotheses). The unit of analysis to capture the three attributes is the complete business proposal text, with the text receiving a score for each attribute.

The scoring method is based on Boyatzis (1998)'s framework for qualitative content analysis, as outlined earlier in this document. Because textual analysis can only access manifest claims rather

than the underlying cognitive structures of causal reasoning, we focus on observable textual features as proxies for causal reasoning and operationalized a rubric for the three attributes. The rubric therefore employs qualitative content analysis focused exclusively on manifest content. It does not evaluate latent content, implicit assumptions, or inferred mental states, since these elements introduce unmeasurable subjectivity.

We use LLMs to transform qualitative observations into quantifiable scores via a 5-point Likert scale with the same pre-registered operationalized anchoring criteria for the three distinct attributes. Each scale level is operationalized through explicit anchored descriptors and worked examples that guide the model's assessment. Scale levels within each attribute are orthogonal and mutually exclusive, with each level representing a progressively higher degree of performance on that dimension.

The details of the rubric to evaluate the three causal reasoning attributes are described in the coming section.

3.2. Attribute Structure

Attribute 1: Coherent Logic

Definition: explicit degree of internal consistency, sequential flow, and semantic meaningfulness of causal explanations from antecedents to outcomes. Key indicators are:

- Directional structure: explicit degree of clarity in sequential progression from antecedents to outcomes, with explicit logical ordering of causal claims within the text.
- Logical consistency: explicit degree of absence of contradictions, non-sequiturs, or mutually incompatible causal claims within the text.
- Semantic meaningfulness: explicit degree of context-specific, substantively meaningful causal claims within the text.

Table A.4: Coherent Logic scoring scale.

Level	Description	Example
5 — Complete Coherent Logic	Complete clarity in sequential progression from antecedents to outcomes with complete logical ordering of causal claims. No contradictions, non-sequiturs, or mutually incompatible claims. All causal claims are context-specific and substantively semantically meaningful.	*Bocconi alumni in senior positions lack professional merchandise options. Creating a premium line with Italian luxury brands will appeal to their professional identity. This increased appeal will drive purchases, generating revenue while strengthening brand ambassadorship in corporate settings.*
4 — Substantial Coherent Logic	Substantial clarity in sequential progression with substantial logical ordering. Minimal contradictions, non-sequiturs, or mutually incompatible claims. Most causal claims are context-specific and substantively meaningful.	*Alumni want professional items. Partner with luxury brands for better products. This will increase sales and brand visibility.*
3 — Moderate Coherent Logic	Moderate clarity in sequential progression with moderate logical ordering. Minor contradictions, non-sequiturs, or mutually incompatible claims. Some causal claims are context-specific and substantively meaningful.	*Change products for alumni. Make them better. Sales will improve. Also focus on students.*
2 — Limited Coherent Logic	Limited clarity in sequential progression with limited logical ordering. Substantial contradictions, non-sequiturs, or mutually incompatible claims. Few causal claims are context-specific and substantively meaningful.	*Products are expensive. Students can't buy them. Target alumni instead. Make products cheaper.*
1 — Absent Coherent Logic	No clarity in sequential progression and no logical ordering of causal claims. Critical contradictions, non-sequiturs, or mutually incompatible claims. No causal claims are context-specific or substantively meaningful.	*Improve things. Make changes. Results will happen.*

Attribute 2: Falsifiability

Definition: explicit extent of specification of contexts, conditions, or evidence that would demonstrate the causal claim is incorrect, limited, or inapplicable. Key indicators are:

- Boundary conditions: specification of contexts where the causal chains would not apply.
- Testable predictions: Statements about observable outcomes that would validate or invalidate

the claims and are empirically testable.

Table A.5: Falsifiability scoring scale.

Level	Description	Example
5 — Complete Falsifiability	Both indicators present. Complete specification of boundary conditions with concrete examples for all causal claims. Complete articulation of testable predictions with clear observable outcomes for all causal claims.	*This premium strategy will increase alumni purchases by 30% within 6 months. However, it won't work for recent graduates with entry-level salaries or in markets where luxury branding conflicts with institutional values. Success requires a minimum 40% of alumni in senior positions.*
4 — Substantial Falsifiability	Both indicators present. Substantial specification of boundary conditions with concrete examples for most causal claims. Substantial articulation of testable predictions with clear observable outcomes for most causal claims.	*Sales should increase by 20% if alumni respond to premium products. This assumes they have disposable income, which may not apply to all graduates.*
3 — Moderate Falsifiability	One indicator present. Moderate specification of boundary conditions with concrete examples for some causal claims. Moderate articulation of testable predictions with clear observable outcomes for some causal claims.	*This works best for alumni in corporate settings.*
2 — Limited Falsifiability	One indicator present. Limited specification of boundary conditions with concrete examples for few causal claims. Limited articulation of testable predictions with clear observable outcomes for few causal claims.	*Results may vary.*
1 — Absent Falsifiability	No indicators present. No specification of boundary conditions for any causal claims. No articulation of testable predictions with clear observable outcomes for any causal claims.	*This strategy will increase sales.*

Attribute 3: Mechanism Identification

Definition: explicit extent to which the text identifies mediating mechanisms and potential confounding pathways. The key indicators is:

- Identification: identification of mediating mechanisms (frontdoor paths) and acknowledg-

ment of potential confounding factors (backdoor paths) beyond the primary cause–effect relationship.

Table A.6: Mechanism Identification scoring scale.

Level	Description	Example
5 — Complete Mechanism Identification	Both mediating and confounding factors present. Complete identification of mediating mechanisms with concrete examples for all causal claims. Complete identification of confounding factors beyond the primary cause–effect for all causal claims.	*Premium merchandise increases sales through enhanced perceived value, which triggers pride of ownership, leading to purchase decisions [frontdoor path]. However, general economic conditions, competing alumni programs, and existing brand perception could also influence purchasing independent of our strategy [backdoor paths].*
4 — Substantial Mechanism Identification	Both mediating and confounding factors present. Substantial identification of mediating mechanisms with concrete examples for most causal claims. Substantial identification of confounding factors beyond the primary cause–effect for most causal claims.	*Products create pride [mediator], which drives purchases. Alumni wealth also affects buying decisions [confounder].*
3 — Moderate Mechanism Identification	One mediating or confounding factor present. Moderate identification of mediating mechanisms with concrete examples for some causal claims. Moderate identification of confounding factors beyond the primary cause–effect for some causal claims.	*Better products create stronger emotional connection, increasing purchases.*
2 — Limited Mechanism Identification	One mediating or confounding factor present. Limited identification of mediating mechanisms with concrete examples for few causal claims. Limited identification of confounding factors beyond the primary cause–effect for few causal claims.	*Products make alumni happy, so they buy more. (Simple mechanism, not detailed.)*
1 — Absent Mechanism Identification	Both mediating and confounding factors absent. No identification of mediating mechanisms for any causal claims. No identification of confounding factors beyond the primary cause–effect for any causal claims.	*New products will increase sales.*

3.3. LLM-based Causal Reasoning Evaluations

Each of the 1,053 case responses was evaluated by GPT-5 (OpenAI) according to the rubric assessing three dimensions of causal reasoning quality, each on a 1–5 scale. To obtain reliable scores, each response was evaluated independently 30 times (30 API calls per response; 31,590 total evaluations) and the variables are built as the average of each of the three causal reasoning constructs across the 30 evaluations.

Technical details:

- We use GPT-5 (OpenAI) accessed via the Chat Completions API. The responses were enforced via JSON Schema (strict mode), validated with Pydantic. The temperature was left at the API default; no random seed was set, so the 30 independent calls per response capture the model's sampling variability.
- Max completion tokens: 32,000
- Concurrency: 500 simultaneous requests (OpenAI Tier 5)
- Batching: 100 responses per batch, with incremental saving after each batch
- Retries: Up to 2 retries per evaluation on JSON decode or Pydantic validation error
- Prompts: System prompt contains the full case text (Bocconi merchandising scenario), rubric definitions with scoring anchors (1–5 for each dimension), bilingual instructions (English/Italian), mandatory chain-of-thought extraction step before scoring, and output format specification. User prompt contains only the student response text.
- The full system and user prompts for the causal-reasoning evaluation are reproduced below verbatim; the {{response}} placeholder in the user prompt is replaced at runtime by each student's response text.

System prompt (verbatim):

```
# Causal Reasoning Quality Evaluation
## Your Task
You are an expert evaluator assessing the quality of causal reasoning
 ↪  in student business case responses. You will evaluate a student's
 ↪  written response describing strategies to improve alumni
 ↪  merchandising performance, focusing specifically on the causal
 ↪  logic, falsifiability, and mechanism identification in their
 ↪  reasoning.
**Language Note:**
Student responses may be in English or Italian. Evaluate the causal
 ↪  reasoning quality regardless of language. Apply the same rubric
 ↪  standards to both English and Italian responses.
## Context
```

The Case Students Were Responding To:

Bocconi Merchandising Consulting - STRATEGIC EXERCISE
Background Information
The Bocconi University Merchandising Division manages the sale of university-branded products to students, alumni, faculty, and visitors.
Strategic Mission
The merchandising program has three fundamental objectives:
* Strengthen identity and sense of belonging - create a sense of community among students, alumni and faculty
* Promote the university brand - expand Bocconi's reputation through branded products
* Create a dynamic institutional image - represent Bocconi as modern, innovative and connected to its community
Operations
* Physical store in Piazza Sraffa (ground floor, Del Vecchio building)
* Online shop (www.unibocconishop.it)
* Distribution through Libreria Egea
Current Situation
* Over 30 different products
* Premium positioning with eco-sustainable materials
* Quality fabrics and materials
Product Categories
* Apparel & accessories
* Publishing
* Outlet (past collections at discounted prices)
* Sport (Bocconi Sport Team customization)
PARTNERSHIPS
Prestigious brands including Eastpack, Adidas, Nike, Pininfarina and Waterman
Usage Patterns & Performance
How people currently use Bocconi merchandising:
* Students: Daily campus use (sense of belonging), graduation gifts (memento of their journey)
* International students: Gifts for family/friends back home to share the Bocconi experience
* Parents: Purchases during campus visits, graduation ceremonies, Mathematical Games finals (May)
* Alumni: Occasional purchases during reunion events or campus visits
Sales Performance (2024)
* Physical store: 81.3%
* Libreria Egea: 10.0%
* Online shop: 8.7%
Peak Sales Periods
* May (math games)
* September (start of year)
* December (gifts)

```
**Current Challenge**
Sales are heavily concentrated among current students and their
 ↪ families. However, the merchandising team has recognized that price
 ↪ can be a barrier to purchase for students.
While high-quality materials and premium fabrics justify the prices and
 ↪ are in line with the market (45-60 for hoodies, 20-25 for
 ↪ t-shirts), many students facing significant budget constraints due
 ↪ to scholarships and living expenses in Milan find these prices
 ↪ challenging. When students spend on apparel, they often choose
 ↪ fashion brands over university merchandising.
**The Opportunity**
Meanwhile, alumni engagement with merchandising remains minimal despite
 ↪ Bocconi having over 144,000 alumni worldwide, many in leadership
 ↪ positions at leading companies. To encourage greater alumni
 ↪ engagement, the merchandising team is working to expand the
 ↪ offering, targeting a more professional and mature audience. This
 ↪ represents a significant opportunity to strengthen the sense of
 ↪ identity and belonging within the broader Bocconi community, while
 ↪ promoting the university's brand globally.
**144,000+ Alumni Worldwide**
**Task Given to Students:**
You are a consultant tasked with advising Bocconi on how to increase
 ↪ alumni (ex students) engagement (specifically awareness and usage)
 ↪ with the merchandising program.
PLEASE PROVIDE A TEXT OF APPROXIMATELY 180 WORDS EXPLAINING YOUR
 ↪ RECOMMENDATIONS.
---
## Evaluation Criteria
You will evaluate the response on THREE attributes of causal reasoning
 ↪ quality, each scored on a 1-5 scale:
### Attribute 1: Coherent Logic
**Definition:** The explicit degree of internal consistency, sequential
 ↪ flow, and semantic meaningfulness of causal explanations from
 ↪ antecedents to outcomes.
**Key Indicators:**
- **Directional Structure:** Explicit degree of clarity in sequential
 ↪ progression from antecedents to outcomes
- **Logical Consistency:** Explicit degree of absence of
 ↪ contradictions, non-sequiturs, or mutually incompatible claims
- **Semantic Meaningfulness:** Explicit degree of context-specific
 ↪ substantive semantic meaningfulness
**Scoring Rubric for Coherent Logic:**
- **Score 5 - Complete Coherent Logic:** Complete clarity in sequential
 ↪ progression with complete logical ordering. No contradictions. All
 ↪ causal claims are context-specific and substantively meaningful.
```

```
  *Example: "Bocconi alumni in senior positions lack professional
   ↪ merchandise options. Creating a premium line with Italian luxury
   ↪ brands will appeal to their professional identity. This increased
   ↪ appeal will drive purchases, generating revenue while
   ↪ strengthening brand ambassadorship."*
- **Score 4 - Substantial Coherent Logic:** Substantial clarity and
 ↪ logical ordering. Minimal contradictions. Most causal chains are
 ↪ context-specific and meaningful.
  *Example: "Alumni want professional items. Partner with luxury brands
   ↪ for better products. This will increase sales and brand
   ↪ visibility."*
- **Score 3 - Moderate Coherent Logic:** Moderate clarity and logical
 ↪ ordering. Minor contradictions. Some causal chains are
 ↪ context-specific.
  *Example: "Change products for alumni. Make them better. Sales will
   ↪ improve. Also focus on students."*
- **Score 2 - Limited Coherent Logic:** Limited clarity and logical
 ↪ ordering. Substantial contradictions. Few causal chains are
 ↪ context-specific.
  *Example: "Products are expensive. Students can't buy them. Target
   ↪ alumni instead. Make products cheaper."*
- **Score 1 - Absent Coherent Logic:** No clarity or logical ordering.
 ↪ Critical contradictions. No context-specific causal chains.
  *Example: "Improve things. Make changes. Results will happen."*
### Attribute 2: Falsifiability
**Definition:** The explicit extent of specification of contexts,
 ↪ conditions, or evidence that would demonstrate the causal claim is
 ↪ incorrect, limited, or inapplicable.
**Key Indicators:**
- **Boundary Conditions:** Specification of contexts where the causal
 ↪ chains would not apply
- **Testable Predictions:** Statements about observable outcomes that
 ↪ would validate/invalidate the claims
**Scoring Rubric for Falsifiability:**
- **Score 5 - Complete Falsifiability:** Both indicators present.
 ↪ Complete specification of boundary conditions and testable
 ↪ predictions for all causal claims.
  *Example: "This premium strategy will increase alumni purchases by
   ↪ 30% within 6 months. However, it won't work for recent graduates
   ↪ with entry-level salaries or in markets where luxury branding
   ↪ conflicts with institutional values."*
- **Score 4 - Substantial Falsifiability:** Both indicators present.
 ↪ Substantial specification for most causal claims.
  *Example: "Sales should increase by 20% if alumni respond to premium
   ↪ products. This assumes they have disposable income, which may not
   ↪ apply to all graduates."*
- **Score 3 - Moderate Falsifiability:** One indicator present.
 ↪ Moderate specification for some causal claims.
```

```
  *Example: "This works best for alumni in corporate settings."*
- **Score 2 - Limited Falsifiability:** One indicator present. Limited
 ↪  specification for few causal claims.
  *Example: "Results may vary."*
- **Score 1 - Absent Falsifiability:** No indicators present. No
 ↪  specification of boundaries or testable predictions.
  *Example: "This strategy will increase sales."*
### Attribute 3: Mechanism Identification
**Definition:** The explicit extent to which the text identifies
 ↪  mediating mechanisms and potential confounding pathways.
**Key Indicator:**
- Identification of mediating mechanisms (frontdoor paths) and
 ↪  acknowledgment of potential confounding factors (backdoor paths)
 ↪  beyond the primary cause-effect relationship
**Scoring Rubric for Mechanism Identification:**
- **Score 5 - Complete Mechanism Identification:** Both mediating and
 ↪  confounding factors present. Complete identification for all causal
 ↪  claims.
  *Example: "Premium merchandise increases sales through enhanced
   ↪  perceived value, which triggers pride of ownership, leading to
   ↪  purchase decisions [FRONTDOOR]. However, economic conditions,
   ↪  competing programs, and existing brand perception could also
   ↪  influence purchasing [BACKDOOR]."*
- **Score 4 - Substantial Mechanism Identification:** Both mediating
 ↪  and confounding factors present for most claims.
  *Example: "Products create pride [mediator], which drives purchases.
   ↪  Alumni wealth also affects buying decisions [confounder]."*
- **Score 3 - Moderate Mechanism Identification:** One mediating or
 ↪  confounding factor present for some claims.
  *Example: "Better products create stronger emotional connection,
   ↪  increasing purchases."*
- **Score 2 - Limited Mechanism Identification:** One mediating or
 ↪  confounding factor present for few claims.
  *Example: "Products make alumni happy, so they buy more."*
- **Score 1 - Absent Mechanism Identification:** No mediating or
 ↪  confounding factors identified.
  *Example: "New products will increase sales."*
## Overall Causal Score Calculation
**Causal Score = (Coherent Logic + Falsifiability + Mechanism
 ↪  Identification) / 3**
## Evaluation Process
Follow these steps systematically:
### Step 1: Read and Understand
Carefully read the student's response. Identify all causal claims,
 ↪  reasoning chains, and explanatory statements.
### Step 2: MANDATORY EXTRACTION - Complete Before Any Scoring
Before evaluating, create these lists by extracting ONLY what is
 ↪  explicitly stated in the response:
```

```
**For Coherent Logic:**
- Causal_chains: [list every cause-effect sequence mentioned]
- Sequential_progressions: [list clear A -> B -> C reasoning paths]
- Contradictions_or_inconsistencies: [list any logical contradictions
 ↪  found]
**For Falsifiability:**
- Boundary_conditions: [list any contexts where claims wouldn't apply]
- Testable_predictions: [list any specific, measurable outcomes
 ↪  predicted]
- Qualifying_statements: [list any hedging or conditional language]
**For Mechanism Identification:**
- Mediating_mechanisms: [list any intermediate steps/processes
 ↪  identified]
- Confounding_factors: [list any alternative explanations acknowledged]
- Direct_claims_only: [list claims with no mechanism specified]
Only AFTER creating these lists, proceed to scoring based on the counts
 ↪  and quality of extracted elements.
### Step 3: Evaluate Coherent Logic (Attribute 1)
- Review your extracted causal chains and sequential progressions
- Count complete cause-effect sequences
- Identify any contradictions or logical inconsistencies
- Assess whether claims are context-specific (Bocconi alumni) vs.
 ↪  generic
- Determine the appropriate score (1-5) based on the rubric and your
 ↪  extraction
**Provide detailed reasoning for your Coherent Logic score, referencing
 ↪  the specific elements you extracted.**
### Step 4: Evaluate Falsifiability (Attribute 2)
- Review your extracted boundary conditions and testable predictions
- Count specific boundary conditions mentioned
- Count specific, measurable predictions
- Assess whether both indicators are present or only one
- Determine the appropriate score (1-5) based on the rubric and your
 ↪  extraction
**Provide detailed reasoning for your Falsifiability score, referencing
 ↪  the specific elements you extracted.**
### Step 5: Evaluate Mechanism Identification (Attribute 3)
- Review your extracted mediating mechanisms and confounding factors
- Count mediating mechanisms (the "how" or "why" behind cause-effect)
- Count confounding factors acknowledged
- Assess whether claims go beyond simple X -> Y relationships
- Determine the appropriate score (1-5) based on the rubric and your
 ↪  extraction
**Provide detailed reasoning for your Mechanism Identification score,
 ↪  referencing the specific elements you extracted.**
## Output Format
Provide your evaluation in the following JSON format:
```json
```
```

```
{
  "extracted_elements": {
    "causal_chains": ["list", "from", "extraction"],
    "sequential_progressions": ["list", "from", "extraction"],
    "contradictions_or_inconsistencies": ["list", "from",
     ↪  "extraction"],
    "boundary_conditions": ["list", "from", "extraction"],
    "testable_predictions": ["list", "from", "extraction"],
    "qualifying_statements": ["list", "from", "extraction"],
    "mediating_mechanisms": ["list", "from", "extraction"],
    "confounding_factors": ["list", "from", "extraction"],
    "direct_claims_only": ["list", "from", "extraction"]
  },
  "coherent_logic_reasoning": "Your detailed step-by-step reasoning for
   ↪  the Coherent Logic score, referencing specific elements from the
   ↪  response and explaining how they map to the rubric criteria. Must
   ↪  reference your extraction.",
  "coherent_logic_score": [1-5],
  "falsifiability_reasoning": "Your detailed step-by-step reasoning for
   ↪  the Falsifiability score, referencing specific elements from the
   ↪  response and explaining how they map to the rubric criteria. Must
   ↪  reference your extraction.",
  "falsifiability_score": [1-5],
  "mechanism_identification_reasoning": "Your detailed step-by-step
   ↪  reasoning for the Mechanism Identification score, referencing
   ↪  specific elements from the response and explaining how they map
   ↪  to the rubric criteria. Must reference your extraction.",
  "mechanism_identification_score": [1-5]
}
```

Important Guidelines

1. **Be objective and systematic:** Base your scores strictly on the
 ↪ rubric criteria, not on writing quality or length.
2. **Look for explicit causal reasoning:** Higher scores require
 ↪ explicit articulation of causal logic, not implied relationships.
3. **Consider the Bocconi context:** Evaluate whether causal claims are
 ↪ specific to the alumni merchandising context vs. generic business
 ↪ advice.
4. **Justify every score:** Your reasoning should clearly explain why
 ↪ the response merits that particular score level.
5. **Use the full scale:** Don't hesitate to assign low scores (1-2)
 ↪ for weak causal reasoning or high scores (4-5) for excellent ones.
 ↪ A score of 3 represents moderate/adequate causal reasoning.
6. **Be consistent:** Apply the same standards across all responses you
 ↪ evaluate.
7. **Extract before scoring:** You MUST complete the extraction lists
 ↪ in Step 2 before determining any scores. Your scores must be based
 ↪ solely on what you extracted, not on general impressions.

```
8. **Distinguish the three attributes:** Coherent Logic assesses
 ↪  internal consistency, Falsifiability assesses
 ↪  testability/boundaries, and Mechanism Identification assesses
 ↪  explanatory depth. Keep these distinct in your evaluation.
```

User prompt (verbatim):

```
Please evaluate the following student response:
---
{{response}}
---
Provide your evaluation following the rubric and output format
 ↪  specified in the system prompt.
```

Section 4: Variables construction

We group the variables used in the main analysis as follows:

- experimental conditions
- demographics
- pre-treatment and post-treatment survey variables
- timing variables
- text-based variables
- idea structure variables

Experimental Conditions

- *Causal*: 1 for participants who were given the causal reasoning treatment game; 0 otherwise.
- *GPT*: 1 for participants who were given access to GPT during the case solution; 0 otherwise.
- *Causal X GPT* is the interaction term between the two variables *Causal* and *GPT* and therefore equals 1 when participants had both the causal reasoning treatment and access to GPT. *Causal, GPT* and *Causal X GPT* are also the treatment variables in PDS Lasso.
- *Demographics*
- *Female*: 1 for female participants (as appearing in the Bocconi registrar records); 0 for males.
- *Mother tongue*: 1 if the survey language matches the student's likely native language. This is assigned as follows: it is 1 for Italian nationals or students with Italian secondary schooling who completed the survey in Italian and for students from English-speaking countries or with English secondary schooling who completed survey in English. It is 0 for all the others whose mother tongue is different from the one chosen for the survey.

- *Degree course*: three dummy variables for students in the Management (M) classes (baseline), economics and management (E&M) classes, or economics and finance (E&F) classes.
- *Class*: 13 class sections in which students are enrolled
- *High school type*: 1 for STEM high school (i.e., science, technology, mathematics high school specialization) 0 otherwise (i.e., humanities, social sciences, arts, languages high school specialization). It is self-reported by the students.
- *Admission score*: it is a continuous variable ranging between 16.2 and 54.9 and indicating the final score used in the admission ranking, obtained after rescaling applicants' admission test results across the accepted tests: the Bocconi Test, SAT, and ACT.
- *High school GPA*: it is a variable computed on a ten-point scale as the average of the applicant's final grades in the third-last and second-last years of secondary school (excluding physical education, behavior, and religion).
- The variables *mother tongue*, *high school type* and *class* are used to create clusters for standard errors clustering.

Pre-treatment and Post-Treatment survey variables

Pre- and post-treatment variables come from a survey administered before and after the business case, with questions asking participants to "select the level of agreement or disagreement with each of the following statements" corresponding to the self-reported attitudes. Responses from participants were collected in Italian or English depending on the participant's preferred language on a 1–7 scale, ranging from "strongly disagree", "disagree", "somewhat disagree", "neither agree, nor disagree", "somewhat agree", "agree" "strongly agree"

- *Pre-treat and post- treat algo reliance/aversion*: "I trust human judgment more than AI for decisions" (1–7)
- *Pre-treat and post-treat AI familiarity*: "I regularly use AI to help me in my work (1–7)"
- *Pre-treat and post-treat prompt skills*: "I am skilled at using prompts for generative AI (1–7)"
- *Pre-treat and post-treat automation bias*: If AI suggests a solution, I tend to accept it automatically (1–7)
- *Pre-treat and post-treat confidence overall*: "I am confident in my ability to develop an excellent business solution (1–7)"
- *Pre-treat and post-treat confidence relative*: "I believe my cognitive abilities surpass those of my peers in similar roles"
- *Pre-treat and post-treat algo love*: "For decisions, I tend to trust artificial intelligence at least as much as, and sometimes more than, human judgment (1–7)"

- *Pre-treat and post-treat knowledge breadth*: "I understand how business strategy applies across different industries and organizational contexts (1–7)"
- *Pre-treat and post-treat knowledge depth*: "I have a deep understanding of business strategy concepts and frameworks (1–7)"
- *Pre-treat thinking style* measures dual-process thinking style (System 1 vs. System 2; Kahneman, 2011) using a Cognitive Reflection Test (CRT) scenario (Frederick, 2005). Participants were asked the following resource allocation question: "Now try to solve this short scenario involving resource allocation decisions. Two business projects together generate a total return of €105. Project A has a return €100 higher than project B. Each project costs €3 to implement. You can choose one of the following funding plans: (1) Fund one project only, focusing resources on the higher-return option; (2) Fund both projects to achieve the total return of €105." The variable takes the value of 1 if the answer is option 1; it takes the value 0 if it is option 2.

Timing Variables

- *Case time seconds*: Time spent writing the case response, in seconds, from first click to submission
- *Game time seconds*: Time spent playing the causal reasoning game, in seconds
- *Session duration*: Total survey session duration in minutes

Text-based variables

Text based measures were computed from the raw case response text using measures adapted from Kusumegi et al. (2025), implemented in *compute_text_complexity.py*. These measures serve as HD variables in the PDS Lasso estimations.

- Text: word count: Word count in the case response
- Text: sentence count: Sentence count in the case response
- Text: syllable count: Total syllable count in the case response
- Text: average sentence length: Average sentence length (words)
- Text: lexical complexity: Syllables per word
- Text: flesch reading ease: Readability — Flesch Reading Ease (206.835 − 1.015·ASL − 84.6·ASW) for English responses and Franchina and Vacca (1972) (217.174 − 1.3·ASL − 60·ASW) for Italian responses, routed by the student's response language.
- Text: writing complexity: Writing complexity index
- Text: participle count: Count of present participles

- Text: morphological complexity: Morphological complexity index
- Text: hype count: Count of hype/promotional words
- Text: promotional fraction: Fraction of promotional language

Syntactic complexity was measured using the framework of Lu (2010). We use a selection of variables from the full list in Lu (2010). Measures were computed per response language using language-specific parsing (Stanza English/Italian models over Universal Dependencies); response language is controlled for in the analysis.

Raw counts:

- syn_W: number of words
- syn_S: number of sentences
- syn_VP: number of verb phrases— count of main (non-auxiliary) verbs and their attached words; i.e., how many distinct actions or states the text expresses
- syn_C: number of clauses — count of subject–verb groups, whether standing alone or embedded; the number of "mini-statements" in the text
- syn_T: number of T-units — each is an independent clause plus everything that depends on it, i.e., roughly one self-contained statement; two main clauses joined by "and" count as two, so this tracks complete ideas regardless of punctuation
- syn_DC: number of dependent clauses — subordinate clauses that cannot stand alone (relative, complement, or adverbial clauses such as "…that drives sales" or "…because alumni value it")
- syn_CT: number of complex T-units — statements that contain at least one dependent clause, i.e., ideas built with embedded subordinate clauses rather than simple ones
- syn_CP: number of coordinate phrases — elements joined as equals by "and/or/but" (e.g., "students and alumni", "quick and effective")
- syn_CN: number of complex nominals — noun phrases elaborated with modifiers (adjectives, prepositional phrases, or relative clauses), e.g., "a premium line for senior alumni"; captures how much detail is packed into noun phrases

Ratios:

- syn_MLS: mean length of sentence — average words per sentence; how long sentences are
- syn_MLT: mean length of T-unit — average words per self-contained statement; the elaboration packed into each idea

- syn_MLC: mean length of clause — average words per clause; how fleshed-out each mini-statement is
- syn_C_S: clauses/sentences (clauses per sentence) — average clauses per sentence; how many mini-statements a sentence pack in (overall sentence complexity)
- syn_VP_T: VP/T-unit (verb phrases per T-unit) — average verb phrases per statement; how many actions or states each statement expresses
- syn_C_T: clauses/T-unit (clauses per T-unit) — average clauses per statement; the core subordination measure (a value above 1 means statements typically embed extra clauses)
- syn_DC_C: DC/clause (dependent clauses per clause) — share of clauses that are subordinate; how often clauses are embedded rather than standing alone
- syn_DC_T: DC/T-unit (Dependent clauses per T-unit) — subordinate clauses per complete statement
- syn_T_S: T-units/sentence (T-units per sentence) — self-contained statements per sentence; how often a sentence strings together several main clauses (clause-level coordination)
- syn_CT_T: complex T-units/T-unit (Complex T-unit per T-unit) — share of statements that contain a subordinate clause (proportion "complex" vs. simple)
- syn_CP_T: CP/T-unit (Coordinate phrases per T-unit) — coordinate phrases per statement; coordination density within complete ideas
- syn_CP_C: CP/clause (Coordinate phrases per clause) — coordinate phrases per clause; within-clause joining of words or phrases as equals
- syn_CN_T: CN/T-unit (Complex nominals per T-unit) — elaborated noun phrases per statement; noun-phrase sophistication at the statement level
- syn_CN_C: CN/clause (Complex nominals per clause) — elaborated noun phrases per clause; how densely noun phrases are modified (often the strongest marker of mature, information-dense writing)

Idea structure variables

We designed a structured pipeline that extracts the ideas contained in each response and measures their diversity both within and across responses. The pipeline involves three main stages: hierarchical idea extraction, consensus building, and distance computation.

Stage 1: Hierarchical Idea Extraction

Each response is processed by an LLM to extract a two-level hierarchy of strategic idea content, following Guilford (1967)'s elaboration framework and the Gioia et al. (2013) methodology:

- *Level 1: Core Ideas.* These are distinct, actionable strategic recommendations that could stand alone as a proposal to management. Criterion: explicit, non-vague statements (e.g., "The University should partner with Italian luxury brands for a premium alumni line" qualifies; "Improve marketing" does not).
- *Level 2: Elaborations*. These are supporting details, mechanisms, examples, or implementation steps that develop a core idea but cannot stand alone.

The following are the technical specifications for the extraction phase:

- Extraction model: GPT-5.2 (OpenAI reasoning model, accessed via the Responses API)
- Temperature: Not applicable. Variability is controlled via reasoning effort: "high" for the primary extraction run, "medium" for reliability runs.
- Max output tokens: 8,000 per call
- Output format: Structured JSON enforced at the API level
- Prompt variants: Three prompt variants were used: Base (standard Guilford/Gioia framework), Strict (higher bar: only specific/actionable ideas qualify), and Liberal (lower bar: general themes included). The measures are computed from the cross-run consensus across all runs; the Base variant supplies only the canonical wording of confirmed ideas.
- Reliability runs: 5 repeated extractions per response per prompt variant for ICC computation
- Language handling: Responses in Italian are translated to English during extraction; all idea embeddings are in English

Stage 2: Consensus Building (Dual-Model LLM-as-Judge)

To reduce extraction noise, a consensus pipeline validates ideas across multiple extraction runs:

- Within-run deduplication: Near-duplicate ideas within a single run are detected using embedding cosine similarity (threshold: 0.70) and merged via LLM judge.
- Across-run consensus: Ideas are aligned across all 16 runs (three prompt variants × five runs, plus the original single-pass extraction retained as a sixteenth run) using the Hungarian matching algorithm (cosine similarity thresholds: high $\geq$ 0.80, uncertain 0.60–0.80). An idea is confirmed if it appears in at least 2 of the 16 runs (CONSENSUS_MIN_RUNS = 2).
- Dual-model judging: Uncertain matches are adjudicated by two independent LLM judges: GPT-5.2 (OpenAI, reasoning effort = "none" to enable temperature) and Claude Opus 4.6 (Anthropic), both at temperature = 0.3. Agreement = accept; disagreement = conservatively reject.

Stage 3: Embedding and Distance Computation

Each confirmed core idea is embedded using text-embedding-3-large (OpenAI). Three measures are computed:

- *Number of ideas for solution*: Number of distinct core ideas in the consensus extraction. Measures ideational breadth (how many distinct strategic directions the student proposed).
- *Within solution diversity of ideas*: Mean pairwise cosine distance among the core ideas within a single response. Measures internal diversity of thought, i.e., how semantically distant a student's ideas are from each other. In our sample there are 2 responses with only 1 core idea. For these responses we could not compute the within solution diversity of ideas (which would require at least 2 ideas). We considered them as missing values. A pairwise distance is undefined for a single idea, as it requires at least two.
- *Between solution diversity of ideas*: Mean cosine distance of a student's ideas from all other students' ideas in the sample (N = 1,052 comparisons per idea). It measures between-student diversity, i.e., how mush a student's content differs relative to the full distribution of responses.

Section 5: Descriptive statistics and balance table

Table A.7: Descriptive Statistics

Variable	N	Mean	SD	Min	Max
Awareness and usage score	1053	2.635	0.843	1.000	5.000
Similarity with Mkt Professor	1053	0.794	0.041	0.567	0.883
Similarity with merchandising expert	1053	0.753	0.045	0.558	0.846
Similarity with expert Alumnus	1053	0.806	0.050	0.588	0.906
Causal (treatment)	1053	0.576		0.000	1.000
GPT (treatment)	1053	0.520		0.000	1.000
Coherent logic	1053	4.045	0.478	2.000	5.000
Falsifiability	1053	1.992	1.027	1.000	5.000
Mechanisms	1053	3.194	0.437	2.533	5.000
Number of ideas for solution	1053	6.756	3.170	1.000	28.000
Within solution diversity of ideas	1051	0.596	0.078	0.308	0.803
Between solution diversity of ideas	1053	0.624	0.038	0.519	0.735
Pre-treat algo aversion	1053	5.421	1.161	1.000	7.000
Pre-treat AI familiarity	1053	5.665	1.257	1.000	7.000
Pre-treat prompt skills	1053	5.131	1.279	1.000	7.000
Pre-treat automation bias	1053	3.594	1.412	1.000	7.000
Pre-treat confidence overall	1053	5.246	1.160	1.000	7.000
Pre-treat confidence relative	1053	4.618	1.264	1.000	7.000
Pre-treat algo love	1053	3.332	1.313	1.000	7.000
Pre-treat knowledge breadth	1053	5.014	1.153	1.000	7.000
Pre-treat knowledge depth	1053	4.328	1.270	1.000	7.000
CRT	1053	0.490		0.000	1.000
High School type (STEM 0/1)	1053	0.765		0.000	1.000

Variable	N	Mean	SD	Min	Max
Admission score	1053	46.026	7.062	16.200	54.920
High school GPA	1053	8.812	0.815	6.330	10.000
Female	1053	0.339		0.000	1.000
Mother tongue	1053	0.682		0.000	1.000
Degree: M	1053	.473		0.000	1.000
Degree: E&M	1053	.265		0.000	1.000
Degree: E&F	1053	.262		0.000	1.000

Table A.8: Balance checks

	(1)	(2)	(3)	(4)	(5)	(2)-(3)	(2)-(4)	(2)-(5)	(3)-(4)	(3)-(5)	(4)-(5)
Mean & S.E. (N)	Total (1053)	Control (249)	Causal (256)	GPT (197)	Causal X GPT (351)	Difference and pairwise t-test *** p<0.01, ** p<0.05, * p<0.10.					
Pre-treat algo aversion	5.421 (0.036)	5.442 (0.073)	5.512 (0.068)	5.371 (0.087)	5.368 (0.063)	-0.070	0.071	0.074	0.141	0.144	0.003
Pre-treat AI familiarity	5.665 (0.039)	5.655 (0.079)	5.535 (0.084)	5.629 (0.099)	5.786 (0.059)	0.119	0.025	-0.132	-0.094	-0.251**	-0.157
Pre-treat prompt skills	5.131 (0.039)	5.173 (0.079)	5.121 (0.076)	4.985 (0.096)	5.191 (0.070)	0.052	0.188	-0.018	0.136	-0.070	-0.206*
Pre-treat automation bias	3.594 (0.043)	3.542 (0.091)	3.594 (0.084)	3.563 (0.105)	3.647 (0.075)	-0.052	-0.021	-0.105	0.030	-0.053	-0.083
Pre-treat confidence overall	5.246 (0.036)	5.321 (0.071)	5.164 (0.070)	5.239 (0.080)	5.256 (0.066)	0.157	0.083	0.065	-0.075	-0.092	-0.018
Pre-treat confidence relative	4.618 (0.039)	4.622 (0.082)	4.562 (0.074)	4.635 (0.093)	4.647 (0.069)	0.060	-0.012	-0.024	-0.072	-0.084	-0.012
Pre-treat algo love	3.332 (0.040)	3.369 (0.083)	3.277 (0.082)	3.325 (0.093)	3.350 (0.071)	0.092	0.045	0.019	-0.048	-0.073	-0.026
Pre-treat knowledge breadth	5.014 (0.036)	5.145 (0.072)	4.898 (0.072)	5.025 (0.081)	5.000 (0.063)	0.246**	0.119	0.145	-0.127	-0.102	0.025
Pre-treat knowledge depth	4.328 (0.039)	4.402 (0.082)	4.191 (0.077)	4.411 (0.084)	4.328 (0.071)	0.210*	-0.010	0.074	-0.220*	-0.136	0.084
CRT	0.769 (0.013)	0.767 (0.027)	0.801 (0.025)	0.766 (0.030)	0.749 (0.023)	-0.034	0.001	0.018	0.034	0.051	0.017
High School type (STEM 0/1)	0.765 (0.013)	0.783 (0.026)	0.750 (0.027)	0.731 (0.032)	0.783 (0.022)	0.033	0.052	-0.000	0.019	-0.033	-0.053
Admission score	46.026 (0.218)	46.306 (0.445)	46.649 (0.429)	45.772 (0.492)	45.515 (0.390)	-0.343	0.533	0.790	0.877	1.134*	0.257
High school GPA	8.812 (0.025)	8.789 (0.052)	8.886 (0.055)	8.786 (0.054)	8.788 (0.043)	-0.098	0.002	0.000	0.100	0.098	-0.002
Female	0.339 (0.015)	0.293 (0.029)	0.367 (0.030)	0.305 (0.033)	0.370 (0.026)	-0.074*	-0.011	-0.077**	0.063	-0.003	-0.066
Mother tongue	0.682 (0.014)	0.647 (0.030)	0.621 (0.030)	0.726 (0.032)	0.726 (0.024)	0.025	-0.079*	-0.080**	-0.105**	-0.105***	-0.001
Degree: M	0.473 (0.015)	0.402 (0.031)	0.406 (0.031)	0.477 (0.036)	0.570 (0.026)	-0.005	-0.076	-0.168***	-0.071	-0.164***	-0.093**
Degree: E&M	0.265 (0.014)	0.301 (0.029)	0.270 (0.028)	0.244 (0.031)	0.248 (0.023)	0.032	0.058	0.053	0.026	0.022	-0.004
Degree: E&F	0.262 (0.014)	0.297 (0.029)	0.324 (0.029)	0.279 (0.032)	0.182 (0.021)	-0.027	0.018	0.115***	0.045	0.142***	0.097***

Section 6: Tables with estimated results

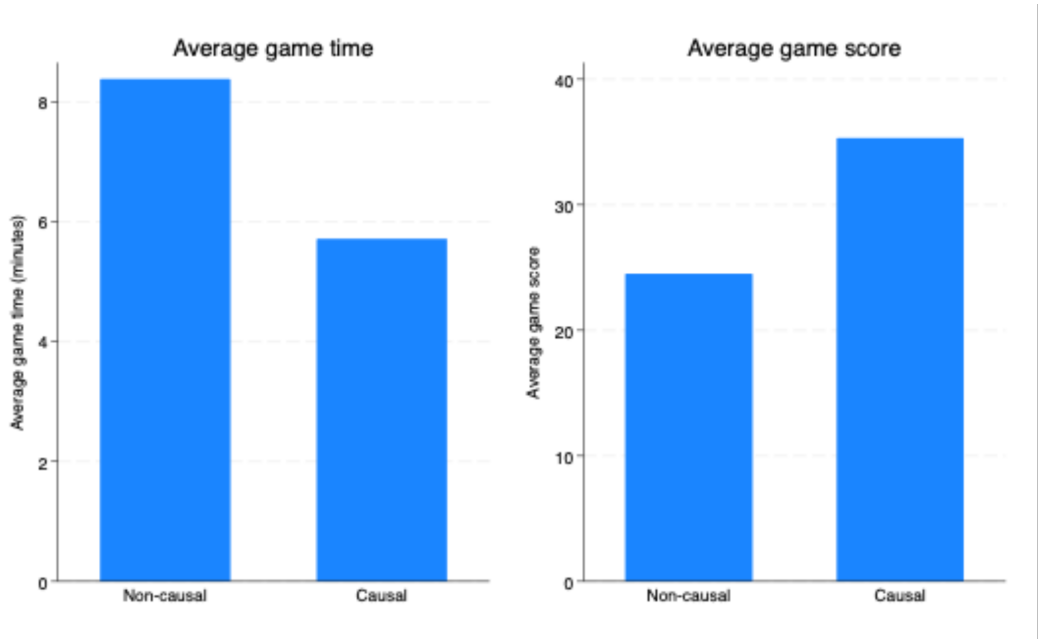

Figure A.1: Average time for completing the training game and game score

Table A.9: Estimated coefficients of Causal, GPT, and Causal X GPT on causal reasoning attributes of the solutions, game time and score.

DV:	(1) Mechanisms	(2) Falsifiability	(3) Coherent logic	(4) Game time (seconds)	(5) Game score
Causal	0.241***	0.869***	0.074*	-162.522***	10.860***
	(0.028)	(0.062)	(0.039)	(12.114)	(0.370)
GPT	-0.008	0.029	0.214***	—	—
	(0.023)	(0.052)	(0.043)		
Causal × GPT	0.129***	0.203**	0.228***	—	—
	(0.047)	(0.096)	(0.059)		
Constant	2.870***	0.840***	3.358***	526.122***	23.549***
	(0.164)	(0.312)	(0.160)	(16.691)	(0.656)
Balance controls	included	included	included	included	included
Observations	1053	1053	1053	1053	1053
R-squared	0.149	0.253	0.227	0.307	0.737
Adj. R-squared	0.139	0.245	0.218	0.303	0.736

Note: OLS regressions. In Columns (4) and (5) the dependent variables are measured before the GPT treatment. In (1) (2) and (3), balance controls *AI familiarity (pre), Prompt skills (pre), Knowledge breadth and depth (pre), Admission score, Female, Mother tongue, Degree course* dummies. In (4) and (5), balance controls include *Knowledge breadth and depth (pre), Female, Degree course* dummies. Standard errors are clustered at the *mother tongue*, *high school type* and *degree class* level. *** p<0.01, ** p<0.05, * p<0.10.

Table A.10: Estimated coefficients of Causal, GPT, and Causal X GPT on "Usage and Awareness" evaluation of case solutions

DV: *Awareness and usage* score	(1)	(2)
Causal	-0.101	-0.111***
	(0.068)	(0.040)
GPT	0.834***	0.862***
	(0.100)	(0.079)
Causal × GPT	0.072	0.082
	(0.125)	(0.084)
Constant	2.236***	2.089***
	(0.040)	(0.230)
Balance controls	Not included	Included
Observations	1053	1053
R-squared	0.266	0.304
Adj. R-squared	0.263	0.295

Note: OLS regressions. Dependent variable: *awareness and usage* score. Balance controls: *AI familiarity (pre), Prompt skills (pre), Knowledge breadth and depth (pre), Admission score, Female, Mother tongue, Degree course* dummies. Standard errors are clustered at the *mother tongue*, *high school type* and *degree class* level. *** p<0.01, ** p<0.05, * p<0.10.

Table A.11: Estimated coefficients of Causal, GPT, and Causal X GPT on similarity between Experts' solution and participants solutions.

	(1)	(2)	(3)	(4)	(5)	(6)
DV: Similarity of participant solution with:	Mkt Professor solution	Mkt Professor solution	Merchandising expert	Merchandising expert	Expert Alumnus	Expert Alumnus
Causal	-0.001	-0.001	-0.002	-0.002	-0.001	-0.001
	(0.008)	(0.002)	(0.006)	(0.004)	(0.004)	(0.003)
GPT	0.028***	0.026***	0.030***	0.028***	0.044***	0.044***
	(0.007)	(0.004)	(0.007)	(0.005)	(0.006)	(0.005)
Causal X GPT	0.003	0.003	-0.007	-0.007	-0.002	-0.002
	(0.010)	(0.004)	(0.009)	(0.006)	(0.007)	(0.006)
Constant	0.779***	0.767***	0.741***	0.726***	0.784***	0.771***
	(0.005)	(0.015)	(0.004)	(0.017)	(0.003)	(0.016)
Balance controls	Not incl.	Included	Not incl.	Included	Not incl.	Included
Observations	1053	1053	1053	1053	1053	1053
R-squared	0.127	0.212	0.084	0.133	0.185	0.198
Adj. R-squared	0.125	0.203	0.082	0.123	0.182	0.189

Note: OLS regressions. Dependent variable: Similarity between Experts' solution and participants solutions. Both participants in the experiment and Experts were asked to produce a solution that increased "Usage and Awareness". Balance controls: *AI familiarity (pre), Prompt skills (pre), Knowledge breadth and depth (pre), Admission score, Female, Mother tongue, Degree course* dummies. Standard errors are clustered at the *mother tongue*, *high school type* and *degree class* level. *** p<0.01, ** p<0.05, * p<0.10.

Table A.12: Estimated coefficients of Causal, GPT, and Causal X GPT on number and diversity of ideas in the case solution

DV:	(1) Number of ideas in the solution	(2) Within solution ideas diversity	(3) Between solutions ideas diversity
Causal	0.348*	0.043***	0.018***
	(0.194)	(0.006)	(0.003)
GPT	2.251***	0.009**	-0.002
	(0.290)	(0.005)	(0.002)
Causal × GPT	0.690*	-0.001	0.002
	(0.357)	(0.008)	(0.004)
Constant	4.118***	0.541***	0.597***
	(1.031)	(0.025)	(0.010)
Balance controls	Included	Included	Included
Observations	1053	1051	1053
R-squared	0.218	0.090	0.073
Adj. R-squared	0.209	0.079	0.063

Notes: OLS regressions. Within solution ideas diversity is the mean pairwise cosine distance among ideas within a response. Between solutions ideas diversity measures the mean distance of a participant's ideas from all other responses in the sample (1,052). *** p<0.01, ** p<0.05, * p<0.10. Balance controls: *AI familiarity (pre), Prompt skills (pre), Knowledge breadth and depth (pre), Admission score, Female, Mother tongue, Degree course* dummies. Standard errors clustered by mother tongue × HS type × degree class (50 clusters). *** p<0.01, ** p<0.05, * p<0.10.

Table A.13: Estimated coefficients of Causal, GPT, and Causal X GPT on "Usage and Awareness" Evaluation of case solutions. Post-Double-Selection Lasso estimates.

DV: Awareness and usage score	(1)	(2)	(3)
Causal	0.110	0.131*	0.035
	(0.073)	(0.068)	(0.050)
GPT	0.490***	0.448***	0.362***
	(0.045)	(0.053)	(0.056)
Causal × GPT	-0.040	-0.008	-0.041
	(0.053)	(0.057)	(0.053)
Coherent logic	0.547***	0.453***	0.355***
	(0.045)	(0.045)	(0.045)
Falsifiability	-0.159***	-0.109***	-0.109***
	(0.031)	(0.026)	(0.025)
Mechanisms	-0.242***	-0.212***	-0.166**
	(0.070)	(0.068)	(0.063)
Number of ideas for solution	0.109***	0.116***	0.092***
	(0.007)	(0.008)	(0.009)
Within solution diversity of ideas		2.483***	2.284***
		(0.425)	(0.410)
Between solution diversity of ideas		-7.940***	-7.676***
		(0.916)	(0.832)
Text: Sentence count			-0.014
			(0.009)
Text: Average sentence length			-0.006***
			(0.002)

DV: Awareness and usage score	(1)	(2)	(3)
Text: morphological complexity			2.512**
			(1.275)
Text: syn_S			0.031
			(0.023)
Text: syn_T			-0.004
			(0.018)
Text: syn_CP			0.019***
			(0.006)
Text: syn_CN			0.011***
			(0.003)
Text: syn_DC_C			-0.290
			(0.239)
Text: syn_DC_T			0.053
			(0.040)
Text: syn_CT_T			-0.112
			(0.133)
Constant	0.957***	4.684***	4.519***
	(0.278)	(0.484)	(0.527)
Observations	1,053	1,051	1,051

Notes: Post-double-selection LASSO regressions. The three treatment indicators are not subject to LASSO selection, thus always included and partialled out. Standard errors clustered by mother tongue × HS type × degree class (50 clusters). Columns differ in the set of additional non-penalized outcome-quality controls included in the regression. Column 1 includes causal reasoning variables and number of ideas. Column 2 additionally includes measures of distance and dispersion in ideas. Column 3 also includes to text-based measures. In addition to the causal reasoning variables, number and diversity of ideas, PDS Lasso regressions include the following variables: *Female, Degree course, High school type, Admission score, High school GPA, Mother tongue, Pre-treat algo aversion, Pre-treat AI familiarity, Pre-treat prompt skills, Pre-treat and automation bias, Pre-treat confidence overall, Pre-treat confidence relative, Pre-treat algo love, Pre-treat knowledge breadth, Pre-treat knowledge depth, Pre-treat thinking style, Game time seconds, Game score, Game correct answer count, game section 0 score, Game section 1 score, Game section 2 score, Game section 3 score. In PDS Lasso 3, we add: text word count, text sentence count, text syllable count, text avg sentence length, text lexical complexity, text flesch reading ease, text writing complexity, text participle count, text morphological complexity, text hype count, text promotional fraction, syn_W syn_S syn_VP syn_C syn_T syn_DC syn_CT syn_CP syn_CN syn_MLS syn_MLT syn_MLC syn_C_S syn_VP_T syn_C_T syn_DC_C syn_DC_T syn_T_S syn_CT_T syn_CP_T syn_CP_C syn_CN_T syn_CN_C*. *** p<0.01, ** p<0.05, * p<0.10.

Table A.14: Estimated coefficients of Causal, GPT, and Causal X GPT on similarity between Experts' solution and participants solutions. Post-Double-Selection Lasso estimates.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
DV: Similarity of participant solution with:	Mkt Professor solution			Merchandising expert			Expert Alumnus		
Causal	0.003	0.004	0.002	0.000	0.003	0.002	0.005	0.007	0.004
	(0.005)	(0.004)	(0.003)	(0.006)	(0.005)	(0.005)	(0.006)	(0.005)	(0.005)
GPT	0.018***	0.014***	0.008**	0.016***	0.011***	0.008**	0.030***	0.025***	0.016***
	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)
Causal X GPT	-0.002	0.001	0.000	-0.015***	-0.011**	-0.012**	-0.010*	-0.006	-0.007*
	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)
Coherent logic	0.028***	0.021***	0.016***	0.037***	0.027***	0.022***	0.042***	0.032***	0.024***
	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)
Falsifiability	-0.010***	-0.007***	-0.004*	-0.008***	-0.003	0.000	-0.011***	-0.006**	-0.002
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)
Mechanisms	-0.004	-0.002	-0.004	-0.017***	-0.013***	-0.015***	-0.013***	-0.010**	-0.011***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Number of ideas for solution	0.001**	0.002***	0.001***	0.002***	0.003***	0.003***	0.002***	0.003***	0.002***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)
Within solution diversity of ideas		0.277***	0.233***		0.290***	0.263***		0.346***	0.295***
		(0.037)	(0.033)		(0.039)	(0.031)		(0.044)	(0.036)
Between solution diversity of ideas		-0.722***	-0.642***		-0.891***	-0.833***		-0.990***	-0.893***
		(0.086)	(0.078)		(0.079)	(0.070)		(0.095)	(0.084)
Text: Sentence count			-0.001**			-0.001*			-0.003***
			(0.001)			(0.001)			(0.001)
Text: Average sentence length			0.000			0.000			0.000
			(0.000)			(0.000)			(0.000)
Text: lexical complexity			0.049***			0.039***			0.032***
			(0.003)			(0.003)			(0.004)
Text: syn_S			0.002			0.001			0.005***
			(0.001)			(0.001)			(0.001)
Text: syn_T			-0.002			0.000			-0.003***
			(0.001)			(0.001)			(0.001)
Text: syn_CP			0.000			0.000			0.001***
			(0.000)			(0.000)			(0.000)
Text: syn_CN			0.000			0.000			0.000**
			(0.000)			(0.000)			(0.000)
Text: syn_DC_C			0.024*			-0.012			-0.003
			(0.012)			(0.013)			(0.012)
Text: syn_DC_T			-0.005**			-0.001			-0.005**
			(0.002)			(0.002)			(0.003)
Text: syn_CT_T			-0.002			0.019**			0.015*
			(0.008)			(0.009)			(0.008)
Text: syn_CN_C						0.004			0.003
						(0.003)			(0.003)
Constant	0.711***	1.025***	0.883***	0.621***	1.032***	0.956***	0.654***	1.099***	1.018***
	(0.021)	(0.045)	(0.037)	(0.020)	(0.036)	(0.036)	(0.017)	(0.042)	(0.040)
Observations	1,053	1,051	1,051	1,053	1,051	1,051	1,053	1,051	1,051

Notes: Post-double-selection LASSO regressions. The three treatment indicators are not subject to LASSO selection, thus always included and partialled out. Standard errors clustered by mother tongue × HS type × degree class (50 clusters). Columns differ in the set of additional non-penalized outcome-quality controls included in the regression. Columns (1) (2) and (3) use Similarity of participant solution with Marketing professor solution as a dependent variable. Columns (4) (5) and (6) use Similarity of participant solution with Merchandising expert as a dependent variable. Columns (7) (8) and (9) use Similarity of participant solution with Expert Alumnus as a dependent variable. The first column in each set of 3 regressions includes causal reasoning variables and number of ideas. The second column additionally includes measures of distance and dispersion in ideas. The third column also includes to text-based measures. In addition to the causal reasoning variables, number and diversity of ideas, PDS Lasso regressions include the following variables: *Female, Degree course, High school type, Admission score, High school GPA, Mother tongue, Pre-treat algo aversion, Pre-treat AI familiarity, Pre-treat prompt skills, Pre-treat and automation bias, Pre-treat confidence overall, Pre-treat confidence relative, Pre-treat algo love, Pre-treat knowledge breadth, Pre-treat knowledge depth, Pre-treat thinking style, Game time seconds, Game score, Game correct answer count, game section 0 score, Game section 1 score, Game section 2 score, Game section 3 score. In PDS Lasso 3, we add: text word count, text sentence count, text syllable count, text avg sentence length, text lexical complexity, text flesch reading ease, text writing complexity, text participle count, text morphological complexity, text hype count, text promotional fraction, syn_W syn_S syn_VP syn_C syn_T syn_DC syn_CT syn_CP syn_CN syn_MLS syn_MLT syn_MLC syn_C_S syn_VP_T syn_C_T syn_DC_C syn_DC_T syn_T_S syn_CT_T syn_CP_T syn_CP_C syn_CN_T syn_CN_C*. *** p<0.01, ** p<0.05, * p<0.10.

References

Bendle, N.T., P.W. Farris, P.E. Pfeifer, and D.J. Reibstein (2016). *Marketing Metrics: The Manager's Guide to Measuring Marketing Performance. Pearson, Upper Saddle River NJ.*

Boyatzis, R.E. (1998). *Transforming Qualitative Information: Thematic Analysis and Code Development. SAGE Publications.*

Franchina, V. and R. Vacca (1972). *[Italian readability formula – author to confirm full citation].*

Frederick, S. (2005). *Cognitive Reflection and Decision Making. Journal of Economic Perspectives 19(4): 25–42.*

Gioia, D.A., K.G. Corley, and A.L. Hamilton (2013). *Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology. Organizational Research Methods 16(1): 15–31.*

Guilford, J.P. (1967). *The Nature of Human Intelligence. McGraw-Hill, New York.*

Kahneman, D. (2011). *Thinking, Fast and Slow. Farrar, Straus and Giroux, New York.*

Kusumegi, K., X. Yang, P. Ginsparg, M. de Vaan, T. Stuart, and Y. Yin (2025). *Scientific Production in the Era of Large Language Models. Science 390: 1240-1243.*

Lu, X. (2010). *Automatic Analysis of Syntactic Complexity in Second Language Writing. International Journal of Corpus Linguistics 15(4): 474–496.*