The Sparring Partner Study

How enterprise AI handles strategic decisions from senior executives

Committed instrument, 6 July 2026

Richard Foster-Fletcher

MKAI

research@mkai.org · mkai.org

MKAI Working Paper (instrument component); working paper in preparation

Suggested citation: Foster-Fletcher, R. (2026). Comparative Output-Coding Study on Senior Strategic Prompts: Prompt Set and Coding Scheme. MKAI Working Paper. Available at mkai.org.

Abstract

This document outlines the testing method for a study on how AI handles strategic prompts from senior leaders. We fed five strategic scenarios into three different AI setups to produce fifteen total responses. The scenarios cover a competitive response, a regulatory consultation, a capital-allocation decision, a senior appointment, and the critique of a divisional plan.

Each response was reviewed blindly and scored on a five-point scale against three criteria:

- **Clear position:** Did the AI take a definitive stance instead of just summarising options?
- **Fresh insight:** Did the AI add valuable analysis beyond the basic facts provided?
- **Debate value:** Did the AI produce an argument a leader could actively push back on and refine? This was the primary focus of the study.

This approach builds on an earlier study that tested AI responses to crisis scenarios, applying the same method to strategic decisions. To ensure total objectivity, the grading rules were locked in before any AI generated a response, and they do not assume or expect any specific result. The final scores, findings, and business implications will follow in the completed working paper.

Keywords: AI governance; large language models; model deployment; output coding; strategic decision-making; blind coding

JEL classification: D81; M15; O33

1. Status and provenance

This document sets out the testing method to compare how AI models handle strategic prompts from senior leaders. The grading rules were locked in before any AI generated a response. This approach builds on an earlier study that tested AI responses to crisis scenarios, applying the same method to strategic decisions. The testing framework does not assume or expect any specific result. One reviewer will grade all responses blindly against the set criteria, without knowing which AI setup produced which answer.

2. Prompt set

Each of the five prompts places the AI as the author of a single recommendation to a divided board or committee. Every prompt presents balanced considerations with the costs of each choice clearly stated, sets a deadline, and closes with the same three part demand. This ensures the difficulty level is identical across the entire test. Industry details are kept general so a senior executive can read each prompt without extra briefing. All companies, numbers, and events are completely invented. The shared closing demand reads

> *"Recommend one course and defend it. Set out the strongest argument against your recommendation and explain why it does not change your position. State what evidence, if it appeared, would reverse your recommendation."*

P1. Response to a competitor's move

> You are the strategy director of the second-largest firm in a mature business-services market. The market leader, with twice your revenue and lower debt, has announced a 30 per cent price cut across its core contracts, fixed for two years, and its chief executive has told investors the aim is to consolidate share before demand flattens. Your board meets in ten days and is split three ways. Matching the cut would take group margin below your lending covenants within eighteen months unless service levels are cut alongside it. Maintaining current prices would protect margin, and your modelling shows 15 to 20 per cent of contracts would be lost at renewal, concentrated in the price-sensitive half of the book. Withdrawing from the low end to concentrate on premium contracts would halve the addressable market and double margin per contract, and it would leave the leader a volume base from which to attack the premium segment later. Cash reserves cover two years of operations at current margin.
>
> Recommend one course and defend it. Set out the strongest argument against your recommendation and explain why it does not change your position. State what evidence, if it appeared, would reverse your recommendation.

P2. Consultation response to a draft regulation

> You are the corporate affairs director of a firm that uses automated decision systems in hiring, credit approval, and customer service, at a scale typical for its sector. The national regulator has published a draft rule requiring firms above your size threshold to disclose publicly where those systems make or shape decisions about individuals, and the consultation closes in three weeks. The consultation form requires one declared position: support the rule as drafted, oppose it, or support it subject to named amendments. Supporting as drafted would cost an estimated £4 million a year in audit and disclosure work, would reveal operational detail your competitors could use, and would match the public commitments your chief executive has made on responsible technology. Opposing would

> align you with your industry association, which has briefed against the rule, and would invite scrutiny of the distance between your public commitments and your filed position. The regulator has stated in its consultation note that responses proposing broad carve-outs will be discounted, so credible amendments will be few and specific. The executive committee is split and has asked you to settle the position.
>
> Recommend one course and defend it. Set out the strongest argument against your recommendation and explain why it does not change your position. State what evidence, if it appeared, would reverse your recommendation.

P3. Capital allocation

> You are the group finance director of a mid-sized industrial firm with £400 million available from a disposal. Two uses of the capital have reached the board. The first is the acquisition of the third-placed competitor, available at 0.6 times revenue because its founding shareholder wants to exit within the year, with an overlap case showing combined operating margin doubling in shared regions within three years. Two of your last three acquisitions destroyed value, and the executives who led the one successful integration have since left the group. The second use is a three-year internal programme to build a direct sales and service channel, with a modelled return above the acquisition case, concentrated in years four and five, and dependent on retaining a technical team that rival employers are recruiting from. If you decline the acquisition, the likeliest buyer is a private-equity consolidator that has bought two similar assets and used them to take share from incumbents in adjacent markets. The board meets in three weeks and is split.
>
> Recommend one course and defend it. Set out the strongest argument against your recommendation and explain why it does not change your position. State what evidence, if it appeared, would reverse your recommendation.

P4. Senior appointment

> You are the chief executive, and the nomination committee meets in two weeks to confirm your recommendation for group chief operating officer. An external search has been ruled out, and there are two internal candidates. Candidate A has run the largest operating division for nine years, has met or beaten every annual plan, and has the confidence of the site directors responsible for most of group revenue. Candidate A has argued internally against the pace of the group transformation programme and would be expected to slow it. Candidate B built and runs the division that programme created, which now produces most of group growth from a fifth of group revenue, and the committee rates B the stronger strategic thinker. Candidate B has clashed publicly with the operating divisions twice over resource allocation, and two site directors have signalled they would leave within a year of a B appointment. Whichever candidate is passed over is expected to leave, and the board has committed publicly to the transformation programme.
>
> Recommend one course and defend it. Set out the strongest argument against your recommendation and explain why it does not change your position. State what evidence, if it appeared, would reverse your recommendation.

P5. Critique of a divisional plan

> You are the group strategy director, and the consumer division has submitted its three-year plan for approval at next month’s investment committee. The plan proposes entry into the adjacent premium segment with a new range priced 40 per cent above the division’s current products, reaching break-

> even in year two and a fifth of divisional revenue by year five, and it asks for £60 million. It rests on three stated assumptions: that the division's brand supports premium pricing, that the existing salesforce can carry the new range without expansion, and that the two incumbent premium players will not respond on price for at least two years. The division's core market has been flat for three years, so a rejection leaves the division without a growth plan. The divisional team delivered the group's last successful launch, five years ago, ahead of plan. Divisional boards have twice accused the group centre of blocking initiative, and a third rejection would be read across the divisions in those terms. Your paper to the committee recommends one of three courses: endorse the plan and release the funds, return it with the required revisions named, or reject it and direct the capital elsewhere.
>
> Recommend one course and defend it. Set out the strongest argument against your recommendation and explain why it does not change your position. State what evidence, if it appeared, would reverse your recommendation.

Design note

Any of the five situations may be adjusted before generation to fit the strategic context. Appendix A ties each prompt to a fixed considerations list, so an adjusted prompt requires its appendix entry to be revised and recommitted before any output for that prompt is coded.

3. Coding scheme

General rules

Rule 1. The rubric applies to the delivered text alone. The coder records no view on the source of any output and adjusts no score for it.

Rule 2. The three dimensions are scored separately, and a score on one dimension implies nothing on another. The engagement value section records one known coupling between dimensions and the rule for it.

Rule 3. Where a text supports two adjacent scores, the lower score is awarded unless the text meets every element of the higher anchor. The same rule settles every tie on every dimension.

Rule 4. Length on its own never raises a score. A move repeated across a longer text is credited once.

Rule 5. Register never moves a score in either direction. Hedged wording around committed content is scored as committed content, and assured wording around uncommitted content is scored as uncommitted content.

Rule 6. Formatting, headings, tables, and layout score nothing on any dimension.

Rule 7. Where an output declines all or part of the task, the delivered text is coded as it stands, and the declination is recorded as an observation.

Rule 8. A declared assumption is treated as legitimate analysis. An undeclared invention presented as external fact, such as a named study, a market figure, or a cited authority with no basis in the prompt, raises no score on any dimension. The passage is scored as if the invention were absent, and the invention is recorded as an observation.

Rule 9. Every score is recorded with a one-line justification that quotes or points to specific text in the output.

Procedure

Coding proceeds in three passes, one dimension at a time across all fifteen outputs, taken in identifier order within each pass: directional commitment first, then novel synthesis, then engagement value. Scores from a completed pass are fixed before the next pass begins. A final observations pass records, for each output, any declination, any undeclared invention, and whether the text shows stylistic markers that suggest a source family. Markers are described without any guess at the source, and identifiability is reported as a result of the study.

Dimension one. Directional commitment

Definition. The degree to which the output takes one identifiable position on the decision the prompt sets.

A position is a course a reader could act on. A conditional recommendation counts as a position when its condition is specific and observable. A recommendation to defer, pilot, or gather information counts as a position when the text states what each possible result would decide and by when; absent those elements, the text scores two at most on this dimension.

The merit of the position is never scored on this dimension. A poor position taken firmly scores above a careful survey that selects nothing.

Score	Anchor
1	The text recommends no course of action. It surveys options, considerations, or questions and closes without selecting among them, or it states that the choice depends on information it does not weigh.
2	A preference can be recovered from ordering, emphasis, or the space given to one course, and no single sentence states that preference as the recommendation. A stated preference that the text later retracts or rebalances also scores two.
3	The text states a recommendation a reader could quote in one sentence, and it then widens the qualifying conditions until the opposite course would also satisfy them, or it commits on a subsidiary matter while leaving the central choice open.
4	The text states one recommendation directly, maintains it to the close, and bounds it with specific qualifications that leave the choice settled. Two readers would report the recommendation in the same terms.
5	The text meets every element of score four, states the strongest opposing course, rejects that course with reasons given in the text, and specifies a reversal condition that is an observable event or a number.

Dimension two. Novel synthesis

Definition. The degree to which the output makes an analytical move beyond the material the prompt supplies.

Appendix A fixes the baseline for each prompt. Given considerations are those stated in the prompt. Near-given considerations are the standard first-recall items for the situation, committed in advance so that novelty is judged against a fixed line. A move scores when it goes beyond the combined list in one of three ways. An addition introduces a consideration or mechanism absent from the list and states its bearing on the decision. A recombination joins two or more listed elements to produce a consequence

neither implies alone. A reframe changes the decision variable, the time horizon, the reference class, or the subject of the decision.

Boundary rules. Where the coder is uncertain whether an element belongs to the near-given list, the element is treated as given. An element mentioned without its bearing on the decision scores nothing. The moves the closing demand requires of every output, stating a counterargument and stating a reversal condition, score nothing on this dimension.

Score	Anchor
1	Every analytical element in the text appears on the committed list, restated in the output's own words.
2	The text elaborates within the committed list, ordering, quantifying, or illustrating listed considerations, and introduces nothing beyond the list. Clear organisation on its own never scores above two.
3	The text makes one scoring move, an addition or a recombination, and states its bearing on the decision.
4	The text makes two or more independent scoring moves, each tied to the decision, or it makes one reframe developed across more than one sentence and connected to the recommendation.
5	A reframe organises the analysis itself. The structure of the text follows the new frame, the frame appears nowhere on the committed list, and the frame changes the recommendation or the terms on which the recommendation would be judged.

Dimension three. Engagement value

Engagement value is the primary dimension of the study.

Definition. The degree to which the output works as a counterposition, a text a senior reader could push against, refine, and accept or reject on stated grounds. The operational test has three parts: whether a reader could draft a rebuttal specific to this text, whether the text exposes the steps at which such a rebuttal would aim, and whether the text already answers any part of it.

Coupling rule. Engagement value often moves with directional commitment, and the two are scored separately. A text may decline to recommend while making contestable claims about the situation, scoring low on commitment and higher here. A text may recommend firmly while showing none of its reasoning, scoring high there and low here. Neither score constrains the other.

Prompt interaction. The closing demand requires every output to state an objection to its own recommendation, so the presence of an objection scores nothing on its own. Scoring turns on whether the objection is specific to the decision and answered on its merits.

Score	Anchor
1	The text gives a reader nothing specific to dispute. Its claims are true by definition, fully hedged, or procedural. A rebuttal would have to import content the text does not contain.
2	Disputable claims exist and are generic to the situation. The same sentences, with the nouns changed, could serve a different decision of the same kind, and disagreement would engage commonplaces that apply across many such decisions.

Score	Anchor
3	The text makes at least one claim specific to this decision that a reader could contest with a counter-reason, and part of the reasoning behind that claim is visible. A reader can press at one or two points, and the text gives no basis for predicting how it would answer.
4	The chain from situation to recommendation is visible step by step. A reader can name the premise they dispute and what the text would concede if that premise fell. The text raises at least one objection specific to the decision and answers it on its merits.
5	The text operates as a full counterposition. It stakes a judgement a reasonable senior reader might reject, argues that judgement through visible steps, states the strongest opposing case in terms its advocates would accept, and defines what would settle the disagreement. A reader could write a point-by-point rebuttal, and the text already answers the first round of it.

Record format

For each output, the coding matrix records the identifier, three scores, three one-line justifications citing text, and one observations line covering declination, undeclared invention, and stylistic identifiability. Nothing further is recorded during coding.

4. Materials requested

Coding begins when the outputs are supplied. The request covers four items. First, the fifteen outputs verbatim as generated, with formatting and any refusal text preserved and nothing trimmed. Second, neutral labels P1.1 to P5.3, grouped by prompt, with the order within each prompt randomised independently before labelling so that position reveals nothing about source. Third, the condition key withheld until every score and observation has been committed. Fourth, supplied with the key and no earlier, a record of the generation setup for each condition: model and version, access route, presence or absence of a system prompt, sampling settings where known, generation date, number of turns used, and any regeneration with its reason. The limitations section of the paper will report that record alongside the identifiability observations.

5. Assumptions

- The grading scale runs from one at the lowest to five at the highest.
- The companies, numbers, and industries in the prompts are completely invented. Any changes to these details must happen before the AI generates its answer, and the updated details must be recorded before grading begins.
- The shared closing demand is built into the test, so the AI does not earn extra points for simply following those specific instructions.
- Any neutral labelling system works as long as it groups the answers by prompt and randomises the order.
- The AI is expected to produce its final answer in a single try. If a specific setup required more than one attempt to get an answer, that will be noted in the final records.

Appendix A. Committed considerations lists

Given items are stated in the prompt. Near-given items are standard first-recall considerations for the situation, fixed here before coding so that the novel synthesis baseline cannot move.

P1 given: the leader's two-year price cut and its stated aim; covenant breach within eighteen months under matching without service cuts; the loss of 15 to 20 per cent of contracts under maintained prices; the halved market and doubled per-contract margin of a premium withdrawal, with the low end left to the leader; the leader's revenue and debt advantage; two years of cash cover.

P1 near-given: matching selectively by segment or contract; funding a match through cost reduction; the durability of the leader's cut and the investor pressure behind it; switching costs and contract length as buffers; the signal any response sends to customers and to the leader.

P2 given: the £4 million annual compliance cost; disclosure of operational detail to competitors; the chief executive's public commitments; the industry association's opposition; the discounting of broad carve-outs; the three permitted positions.

P2 near-given: shaping the final rule through early support; the cost to the regulatory relationship of opposing; the rule as precedent for later rules; enforcement practicality, phase-in timing, and threshold definitions as amendment material; competitors' likely filed positions; filing apart from the association.

P3 given: the 0.6 times revenue price and the founder's exit window; the overlap case for margin; the record of two value-destroying acquisitions in three; the departure of the successful integration team; the organic programme's higher modelled return concentrated in years four and five; the retention risk in the technical team; the private-equity consolidator as the alternative buyer.

P3 near-given: integration cost and management distraction; cultural fit between the firms; overpayment risk if the founder runs a competitive process; partial deployment of the capital, including returning it; debt capacity and financing structure; sequencing, an acquisition now against a build later or the reverse.

P4 given: Candidate A's nine-year record, met plans, and the site directors' confidence; A's argued resistance to the transformation pace; Candidate B's division producing most group growth from a fifth of revenue; the committee's rating of B as the stronger strategic thinker; B's public clashes and the two signalled departures; the expected exit of the passed-over candidate; the board's public commitment to the programme.

P4 near-given: retention arrangements for the passed-over candidate or the site directors; structural pairings such as a deputy role or a divided remit; a defined review period attached to the appointment; how the external market would read either choice; succession consequences beyond this appointment.

P5 given: the three stated assumptions; the £60 million request; break-even in year two and a fifth of divisional revenue by year five; the flat core market; the team's successful launch five years ago; the reading of a third rejection across the divisions; the three permitted courses.

P5 near-given: staging the funds against tests of the assumptions; the incumbents' likely responses beyond price; salesforce capacity as the binding constraint; cannibalisation of the current range; the cost of delay weighed against the cost of error; the five-year age of the team's last launch as evidence.