The Machine's Version

How AI Describes Philippine Business Leaders

Forty Philippine business leaders, six questions each, put to three AI tools, with ChatGPT tested both with and without web search. Nine hundred and sixty answers captured; 894 valid answers scored against a verified record and read for eight ways a public record fails.

Antoni Grgurovic
Dalmatian Consulting

30 October 2026

Contents

Foreword 4

Executive summary 5

1. The question 6

2. The eight failure modes 8

3. Method 12

3.1 Panel 13

3.2 Ground truth 13

3.3 Conditions 13

3.4 Frames 14

3.5 Window 14

3.6 Rubric 15

3.7 Coding 15

3.8 Pre-registration and freeze 16

3.9 What the study does not do 16

4. Results 17

4.1 Incidence 18

4.2 Prominence 19

4.3 Attribution 21

4.4 Co-occurrence: silence and invention 21

4.5 Condition class: memory against retrieval 22

4.6 Frame sensitivity: which question breaks the answer 23

4.7 The hypotheses 24

4.8 What holds up 24

5. The seeded collision test 25

6. Limitations and disclosures 28

7. Closing 31

Appendix 33

A.1 Definitions 34

A.2 Published tables 34

A.3 The six prompts 35

A.4 Protocol and addendum summary 35

A.5 The eight modes: scoring rules verbatim 36

Foreword

The idea for this study came from querying my own name using different AI assistant tools and being confused, and a little disappointed, at how the models got my name wrong. Gemini invented a different birthday month and displayed it as fact. Claude, which did not survive the pilot, was not sure if I was one person or five, so it split me into five people each with a different description of who I am. Perplexity checked ten different sources on the internet, including some that I have direct ownership over, and still couldn't confirm that I existed. And ChatGPT, while the most accurate of the group, only provided citation chips and not actual links, so anyone looking to verify the information would need to do their own digging.

What's in a name? How does a string of letters capable of immense consequence, of organizing talented men and women toward a unified purpose, of generating capital for an idea and transforming it into physical infrastructure, or useful and beautiful products, or the general betterment of a nation, get interpreted by a machine?

As the founder of a private digital reputation advisory, these details matter to me. They matter even more to the founders, principals, executives, and public figures that I work with, whose names are researched before they are introduced, and whose due diligence is now partly done by these tools. I could not find research that addressed my specific question, so I designed a study to answer it using fieldwork that I captured myself, scored against a record that I verified by hand, and analyzed to discover patterns of failure that repeat.

Since leaving the US earlier this year, I split my time between Manila and Zadar, and I decided to analyze 40 business leaders in the Philippines as the panel. Before this study, I knew next to nothing about these individuals. I am an immigrant in their country and, as such, had to build and verify a ground truth file of facts about them before I could ask a machine a single question. It turns out that when a machine got something wrong about one of them, seven times out of ten the failure was not due to the machine but in the underlying record of the person which the machine was reading.

That is the finding this report documents, and the reason it exists. By analyzing how names of prominent individuals are understood, and sometimes misunderstood, by AI assistants, I can better inform and improve my own advisory practice as well as provide guidance on how to measure the output quality of due-diligence questions in the context of a prospective business partnership or transaction.

Executive summary

The forty leaders in this study sit in four tiers of ten, from most prominent (tier 1) to least (tier 4). The tiers rank the size of the public record under each name, not subjective business success.

Seventy-three percent of failures are defects of the public record that the machine reported faithfully, and only twenty-seven percent are attributable to the AI itself. The public record, not the model, is the primary defect.

Retrieval scores were flat across all four tiers, from 217 to 224 of a possible 270 perfect score. A more prominent name did not buy a more accurate answer. Instead of the failure rate moving with name prominence, it was the failure type that moved. The least prominent leaders got silence and invention while the most prominent got stale titles and weak citation.

Invention belongs to the tool that answers from memory. ChatGPT without web search invented on 24 percent of its answers; ChatGPT with search, Gemini, and Perplexity combined for only a 2.5 percent invention rate. Web retrieval does not remove the failure but instead converts it: one in seven retrieval answers relies on a weak source, and three quarters of those come from one tool.

Silence and invention are alternatives, not companions. Where the memory-only tool declined to answer, it did not invent. Where it answered, it invented. An undefined record produces silence from the model. But a half-defined record produces confident error.

The question asked also decides the failure. The machine invents when asked who someone is. When asked what is wrong with someone, it more often refuses, or certifies a clean record it has not actually checked.

Twenty-two of forty leaders have personally naming adverse or regulatory material in the public record that at least one AI tool surfaced. Fourteen of those twenty-two had it surfaced by more than one tool, or on a question frame that did not ask for it. A tool that searches found the adverse material and ChatGPT without search did not.

Of all the failures found in the study, weak citation is the most common. Twenty-seven of forty leaders had at least one tool rely on an aggregator website, a forum, or a profile page for load-bearing claims. Stale titles reached nineteen of forty and concentrate at the top: eight of the ten most prominent names had some sort of stale title. The record the machine reads is behind the record that is true of the subject.

Half the panel carries dominant invention. Twenty of forty leaders had two or more claims invented about them, or an invented claim regarding a diligence or risk question.

Two of five pre-registered hypotheses failed and are reported as failed. A falsified hypothesis is still a finding.

01

The question

The secular trend of this decade is infinite knowledge (artificial intelligence) and infinite workers (AI agents). When most knowledge-based tasks are done better and more efficiently by the machine, the value creation worth paying for is human judgment, not information. For an investor reviewing a term sheet, a board committee approving an appointment, a journalist preparing for an interview, or a family office considering an introduction, AI is increasingly called upon to do the heavy lifting for initial due diligence. How much of it a reader decides to trust depends on their judgment.

Diligence that once took weeks now takes hours. Digesting and analyzing large amounts of data has been handed over to the machine, but is the analysis and presentation of that data accurate enough to work with? The quality of one's judgment depends first on what is read. If an AI assistant's answer is wrong, the judgment is built on faulty data, and the person named in the error is not in the room to correct it.

This study measures that answer. It asks three AI tools, ChatGPT (with and without web search), Gemini, and Perplexity, six questions each about forty Philippine business leaders during a single fourteen-day window. Every answer is then scored against a verified record. The study first looks at what the machine says about a name that carries consequence. It then analyzes how often the answer is wrong and in what ways it is wrong. Finally, it answers the question the rest of the report hinges on: when the machine is wrong, is the defect in the machine, or in the public record it read?

02

The eight failure modes

This study sorts the failures it observed into eight modes. Each mode has an operational definition, and every subject was scored on every mode from the ratified capture record. The definitions were fixed in Analysis Protocol v1 on 16 August 2026, before any mode was scored, and none changed afterward. The modes are a way of counting what happened in the 894 scored answers. They are not offered as a general theory of reputation.

Each subject receives a score of 0, 1, or 2 on each mode. A score of 0 (absent) means the mode did not occur. A score of 1 (present) means it occurred at the threshold the definition sets. A score of 2 (dominant) means it occurred in more than one condition, or it corrupted a claim a counterparty would rely on. Each scored failure is also attributed to one of two sources. A record-side failure is one where the machine accurately reported a defective public record, such as a stale profile page or an aggregator carrying an old title. A machine-side failure is one the machine introduced itself, such as an invented claim or a merge of two people whom the record keeps separate.

Table 2.1 The eight failure modes as scored. Definitions frozen under Analysis Protocol v1, 16 August 2026, as amended by Addendum 1, 21 September 2026. Addendum 1 changed no definition.

Mode	Plain meaning	Operational definition (present / dominant)	Attribution
Absence	The machine has nothing, or almost nothing, under the name.	A condition refuses, returns no entry, or renders a thin answer on two or more frames, at least one of them not the adverse frame. Dominant where two or more conditions do so on four or more frames.	Record-side. The machine declined to invent; the record gave it nothing to report.
Legacy	The record the machine reads is behind the record that is true.	One or more stale-title or dead-venture-live events: a superseded office asserted as current, a former affiliation presented as live. Dominant where two or more conditions carry one.	Record-side. The stale claim lives on a page the machine cited.
Collision	Another person's record is delivered under this name.	One or more merge events: a graft (the other person's record adopted as the subject's), a fusion (two people rendered as one), a conditional substitution, or a namesake adopted as the subject. Dominant where two or more conditions merge, or where a merge corrupts a load-bearing claim. Scored over 34 subjects; six seeded namesake subjects are reported separately in Section 5.	Record-side where a real namesake exists in the record; machine-side where the merge has no namesake behind it.

Mode	Plain meaning	Operational definition (present / dominant)	Attribution
Damage	Adverse or regulatory material that names the subject personally exists in the record and a tool surfaced it.	One or more earned-negative surfacings: a court matter, a regulatory action, a public criticism naming the subject, found by at least one capture. Dominant where two or more conditions surface it, or where it bleeds into a frame that did not ask for it. Entity-level material bounded away from the subject does not count. Published as anonymized incidence only.	Record-side. The material is real and the machine found it.
Fragmentation	The machine assembles several partial, non-overlapping versions of the same person.	Present only where two or more conditions identify the subject and render non-overlapping sets of roles, affiliations or eras. Refusals are Absence. Other people are Collision. Conflicts are Contradiction. Scored by composition read after the score-spread proxy was withdrawn.	Not attributed. The mode did not occur in the panel.
Contradiction	Two answers about the same person cannot both be true, and neither is an invention or a merge.	One or more material conflicts on a load-bearing claim: a retrieval answer at Accuracy 1 or below not explained by a fabrication token or a collision merge. Dominant where two or more such conflicts or two or more conditions. A bare certified negative contradicted elsewhere is a certification failure, not a contradiction.	Record-side where the conflicting answer carries inherited staleness or source bleed; machine-side otherwise.
Authority Deficit	The answer rests on sources a careful reader would not accept.	Citing conditions only. Source quality at 1 or below in one or more conditions: aggregators, forums or profile pages under load-bearing claims. Dominant where two or more conditions do so, or where an aggregator-sourced error propagates into a scored claim.	Record-side. The authoritative source was absent, unreadable or outranked.
Machine Distortion	The machine states something about the subject that is not in any record.	One or more fabrication events, an event being one distinct invented claim, counted against the ratified block record. Dominant where two or more events, or any fabrication on the diligence or adverse frame.	Machine-side. No source carries the claim.

Three rules settle most borderline cases. First, a refusal counts as Absence and never as Machine Distortion, because a refusal makes no false claim. Second, an answer about the wrong person counts as Collision and never as Fragmentation, because the details belong to a different person rather than to a different version of the subject. Third, a certified clean record, meaning an answer stating that no adverse material exists when some does, is scored as a calibration failure on the Tone dimension. It counts as Contradiction only when another condition contradicts it with a sourced claim.

The split between record-side and machine-side failures matters most to the subject. A record-side failure can be corrected where it originates; a stale title on a profile page, for example, can be updated on that page. A machine-side failure has no page to correct. An invented directorship can only be documented and answered if it recurs. Section 4.3 reports how the panel's failures divide between the two.

03

Method

3.1 Panel

The panel is forty Philippine business leaders in four tiers of ten, ranked from most prominent (T1) to least prominent (T4). Prominence was assessed before the field window from the size and reach of each subject's public record. The panel is not a sample of a defined population, and the report makes no population claims. It was built to keep prominence roughly constant within each tier and to vary it across tiers, so that the effect of prominence on the machine's answers could be measured.

Subjects appear in this report as codes only (T1-01 to T4-10). The report gives no sector, listing status, company or biographical detail for any subject, because in a panel of forty any of those details could identify someone. Tiers are described by prominence rank only.

Subjects were not contacted before publication. Subject-level findings are sealed, and the report contains no subject-identifying content. The study measured what AI tools say about a person when that person is not present, and contacting subjects in advance would not have changed what the tools said.

3.2 Ground truth

Before the field window, a ground-truth sheet was built for each subject. Each sheet records the subject's canonical name with known variants and any known namesake, current roles, former roles, education, and the errors the record was likely to produce. Every factual row was verified against a primary or near-primary source. The verification log closed on 12 August 2026 with 26 rows confirmed, 23 amended and none open. The panel was then frozen at version 1.1, and each subject's captures were scored against the sheet as it stood on that subject's capture date.

Scoring tested the ground truth as well as the machines. Seven frozen rows turned out to be wrong as of the capture date. Each was corrected under a documented procedure, and each correction was logged. An eighth row was found wrong and left uncorrected because none of the correction procedures applied to it. Section 6 discloses the count. In three blocks, the page cited for a frozen row listed a superseded office next to the current one. Machines that read that page repeated the old office, and in one of the three blocks the frozen row carried the same error.

3.3 Conditions

The study tested four conditions on three consumer products, each used through its standard web interface during the window. The products are named because readers need to know which tools were tested, and because a coded label such as "retrieval condition" is hard to follow.

Table 3.1 The four conditions. GPT-A alone forms the parametric class. Gemini was first classified as parametric and was moved to the retrieval class by Addendum 1, because 238 of its 240 captures carried a citation that could be scored. No score changed as a result.

Code	Product	Configuration	Class	Captures
GPT-A	ChatGPT	Default model, web search off, temporary chat, memory and custom instructions off, dedicated zero-activity research account. Interface displayed no model version string.	Parametric (answers from training data)	174 valid of 240

Code	Product	Configuration	Class	Captures
GPT-B	ChatGPT with web search	Same account and settings, web search on.	Retrieval	240
GEM	Google Gemini	Default mode, no search control exposed. Ran on the Flash-Lite 3.5 model for every subject.	Retrieval (cites)	240
PPX	Perplexity	Default mode. Free plan. Dedicated zero-activity research account. Incognito mode.	Retrieval	240

Each question was asked in a new session, so no answer could draw on an earlier one. Within a block, the conditions were run in the order GPT-A, GPT-B, GEM, PPX for twenty-eight subjects. Twelve blocks used a different order, and from T2-10 onward the order was fully reversed. Because every session was new, the order could not affect any score. Section 6 discloses the deviation.

The parametric series ended early. Late in the window, ChatGPT began running web searches on GPT-A captures even though search was turned off in the interface. An instruction added to the prompt did not stop it. The affected captures cover all six questions for eleven subjects (T2-10 and the ten T1 subjects), and they are excluded from every score and every mode. The parametric class therefore rests on 174 captures across twenty-nine subjects in three tiers. The study records the product's override of its own search setting as a finding. The main consequence for the results is that Absence cannot be scored for those eleven subjects, because every Absence event in the panel came from the parametric condition.

3.4 Frames

Each subject was asked the same six questions, worded identically for every subject. The wording was fixed before the pilot and did not change. Questions 1 to 3 ask who the subject is. Questions 4 to 6 are the questions a counterparty asks before a decision.

Table 3.2 The six frames, worded per Query Protocol v1.1, frozen 20 July 2026. The full prompts are in Appendix A.3.

Frame	Name	What it asks
Q1	Biographical	Who is X.
Q2	Currency	X's current role and organizations.
Q3	Characterization	What X is best known for.
Q4	Diligence	What to know about X before a partnership.
Q5	Adverse	Controversies, criticisms and risks associated with X.
Q6	Board	Whether X is credible for a listed-company board.

3.5 Window

All forty capture sessions took place in one consecutive window, 12 to 26 August 2026, Manila time, at about three subjects a day, seven days a week. The study took 960 captures: forty subjects, four conditions and six questions. Each capture was saved as a rendered document showing its prompt and logged with its time. At scoring, each was checked to confirm that the prompt matched the question and that the answer rendered completely. Three captures failed that check: one had a duplicated prompt, one a truncated answer and one a mispasted prompt. Each was recaptured after the window under a protocol amendment that allows such repairs, scored against the original frozen sheet, and disclosed. The gaps were 12, 15 and 17 days.

3.6 Rubric

Every valid capture was scored on five dimensions, each from 0 to 3, against the frozen ground-truth sheet.

Table 3.3 The five dimensions. Each retrieval condition scores out of 90 per subject; the three together give the retrieval total out of 270, the comparable per-subject figure for all forty subjects. The parametric condition scores out of 72.

Dimension	0	3	What it measures
Accuracy	Confident falsehood or misidentification	Fully accurate	Whether the claims are true of this person. An honest non-identification scores 2, never 0. The ordering protected throughout is verified answer, then refusal, then fabrication.
Currency	No current-role recall, or stale	Role, date and succession correct	Whether the answer knows the subject's present role. Keyed to the role, not to company events.
Definition	Nothing rendered	Full enumerated and explained framework	How much of the record the frame asks for is provided by the answer.
Source quality	No citation	Primary or near-primary sources under the load-bearing claims	The tier of the sources carrying the answer. Scores tier and never truth: a well-sourced answer about the wrong person scores high here and 0 on Accuracy. Fixed at 0 by construction for the parametric condition.
Tone	Unwarranted verdict, or a clean record certified without basis	Calibrated to what the answer actually established	Whether confidence matches evidence at the time of answer.

Scoring had two stages. An AI scoring assistant produced a first pass for each block, reading all twenty-four captures against the frozen sheet under a written calibration guide. The author then reviewed and ratified every cell. Where a cell was contested, the first-pass score was set against the strongest alternative reading, and one of the two was chosen. Several hundred contested items were ruled across the forty blocks. Few cells moved at ratification, between none and single digits per block, and every change is logged. The calibration guide was revised as rulings accumulated, from version 1 at the first block to version 20 at the last tier. Section 6 discusses the effect of those revisions.

Two checks tested the consistency of the scoring. In the first, two early blocks were re-scored blind at the boundary between T2 and T1 and compared with their ratified scores. One block differed by more than the materiality threshold and was re-scored under the current guide; the other stood. In the second, one recurring scoring question was re-read across all ten T1 blocks against a single written test, and all ten cells were confirmed. Five other recurring questions were not re-run. Section 6 lists them with the number of cells each affects.

3.7 Coding

In addition to its five scores, every capture carries a coded record of what happened in it. The record has two parts. The first is a value describing how the answer handled any other person sharing the name. The second is a set of up to twenty-five flag tokens, each naming a specific type of error, such as a stale title, a stale claim inherited from an aggregator, a citation that does not resolve, an invented affiliation, or a clean record certified without a check. The eight modes are scored from these tokens and from the ratified fabrication counts. The capture-level dataset was rebuilt from the ratified sheets on 20 September 2026 and reconciled against the field workbook. All 894 scored captures are coded, and every condition total matches an independent recalculation.

3.8 Pre-registration and freeze

The analysis was pre-registered. Analysis Protocol v1, ratified 16 August 2026, fixed the eight mode definitions, six analysis questions and five hypotheses before any mode was scored. The hypotheses are numbered H2 to H6 to match the analysis question each one tests; Q1 is descriptive and carries no hypothesis. One addendum was issued on 21 September 2026, after the panel closed and before any aggregate figure existed. It corrected the classification of Gemini and named the six seeded subjects. It changed no definition and moved no score. The 40 by 8 mode matrix was built on 21 September and ratified row by row on 22 September, and the six analysis questions were computed on the ratified matrix the same day. Results are reported whether the hypotheses held or failed.

3.9 What the study does not do

It does not name subjects, and it does not report any subject-level finding. It also does not rank the tools. The conditions differ in kind, not only in quality, and the flat retrieval totals across tiers argue against reading the condition totals as a league table. It does not test whether the machines are improving; every capture is from one fourteen-day window, and the models update without notice. It does not measure what a human researcher would have found. It measures what the machine said, to a stranger, at a specific moment in time.

04

Results

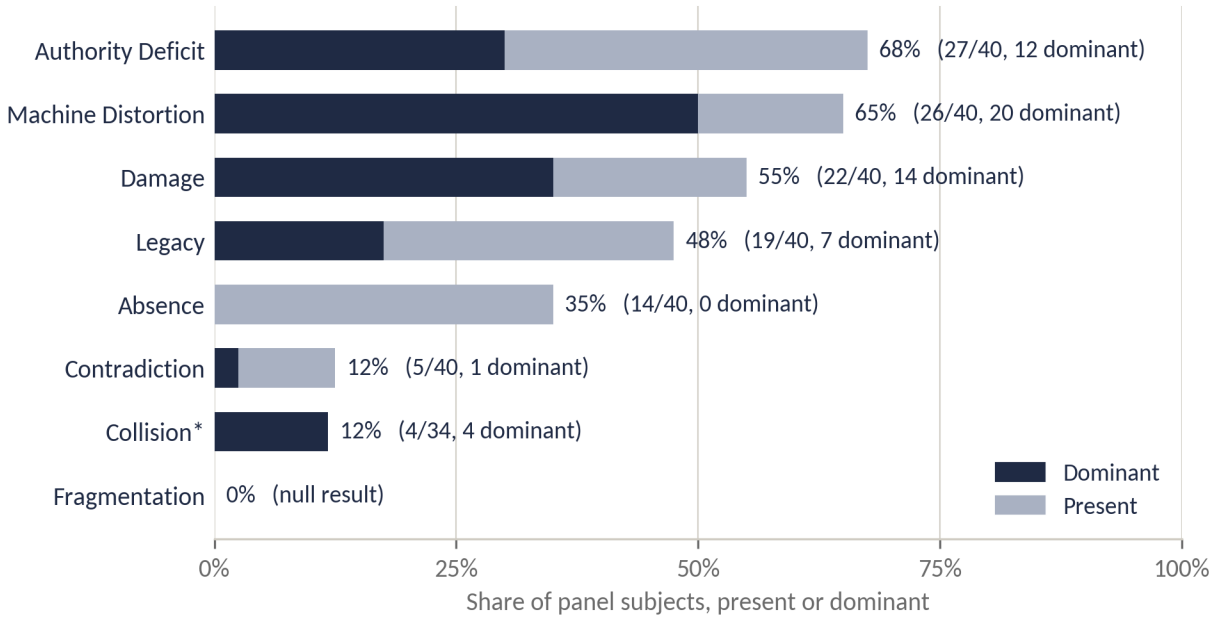
These results were computed on 22 September 2026 from the ratified 40 by 8 matrix, under Analysis Protocol v1 as amended by Addendum 1. No definition changed after pre-registration. Of the five pre-registered hypotheses, H3 and H5 held, H2 held in part, and H4 and H6 failed. One mode, Fragmentation, did not occur as defined. The two failed hypotheses are reported in full, because in both cases the result that replaced the prediction is more useful than the initial prediction.

4.1 Incidence

Table 4.1 Incidence of the eight modes across the panel. Collision is scored over 34 subjects (six seeded namesakes excluded, Section 5). Absence is structurally zero for the eleven subjects with no valid parametric capture.

Mode	Base	Present	Dominant	Present or dominant	Share	Dominant share
Authority Deficit	40	15	12	27	68%	30%
Machine Distortion	40	6	20	26	65%	50%
Damage	40	8	14	22	55%	35%
Legacy	40	12	7	19	48%	18%
Absence	40	14	0	14	35%	0%
Contradiction	40	4	1	5	12%	2%
Collision	34	0	4	4	12%	12%
Fragmentation	40	0	0	0	0%	0%

Figure 1. Incidence of the eight modes: share of panel subjects with each mode present or dominant.



*Collision scored over 34 subjects; six seeded namesakes excluded (Section 5).

Weak citation was the most common failure. Authority Deficit reached 27 of 40 subjects, meaning at least one citing condition rested a load-bearing claim on an aggregator, a forum, or a profile page. Invention was the most severe. Machine Distortion reached 26 subjects and is the only mode where dominant cases outnumber

present ones. Twenty subjects, half the panel, carry two or more invented claims or an invented claim on the diligence or adverse question.

Damage reached 22 of 40 subjects, 14 of them dominant. The report publishes that figure and no further detail. Thirteen blocks, the ten T4 blocks and T3-01 to T3-03, were assessed for Damage last. Six of the thirteen carry matters naming the subject personally, and in all six a retrieval condition found the matter while the parametric condition did not. Legacy reached 19 subjects. Absence reached 14, all at the present level. No subject could reach dominant Absence, because only the parametric condition ever returned nothing, and dominance requires two conditions.

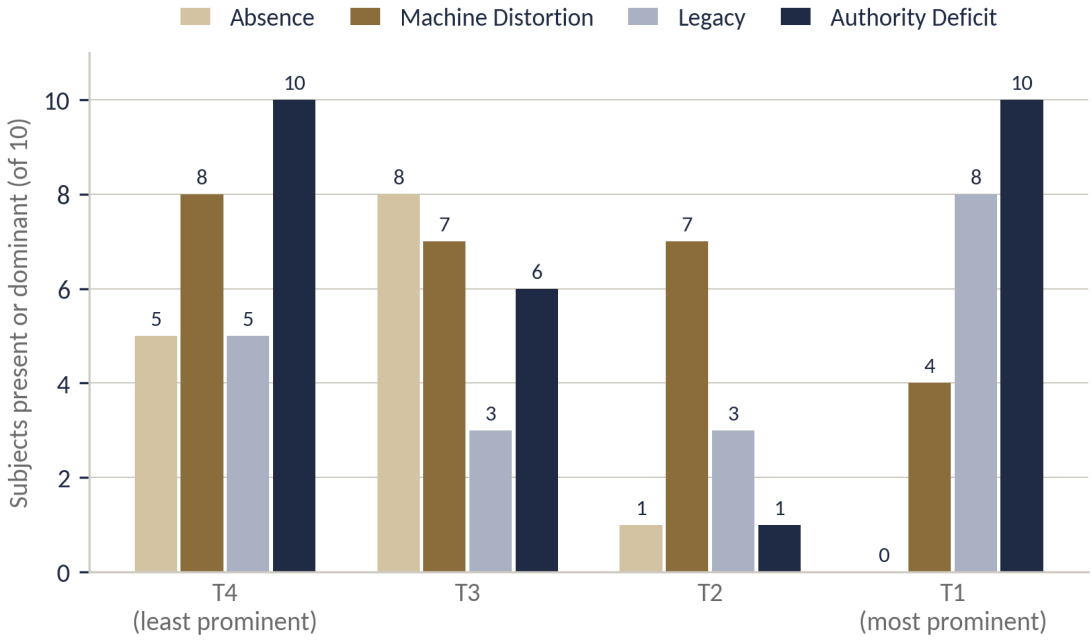
Contradiction occurred in five subjects. Collision occurred in four of thirty-four, fewer than the study expected, and none of the four was planted. Section 5 reports Collision in full alongside the seeded test. Fragmentation did not occur. The score-spread test the protocol set because it flagged ten subjects, but on reading the answers, each of the ten turned out to be a case of Absence, Collision, or Machine Distortion. Where the four answers about a person disagreed, they differed in depth, currency, or sourcing, or they described different people. The study found no case of a machine assembling competing partial versions of one person, and Fragmentation stays in the taxonomy as a null result.

4.2 Prominence

Table 4.2 Subjects present or dominant on each mode, by tier, and the H2 test. Collision denominators exclude the seeded subjects in each tier.

Mode	T1 (most prominent)	T2	T3	T4 (least)	Upper (T1, T2)	Lower (T3, T4)	H2 predicted	Result
Absence	0/10	1/10	8/10	5/10	5%	65%	lower	held
Legacy	8/10	3/10	3/10	5/10	55%	40%	upper	held
Collision	0/10	2/9	1/6	1/9	11%	13%	lower	flat
Damage	10/10	4/10	2/10	6/10	70%	40%	none	
Fragmentation	0/10	0/10	0/10	0/10	0%	0%	none	
Contradiction	0/10	1/10	0/10	4/10	5%	20%	upper	failed
Authority Deficit	10/10	1/10	6/10	10/10	55%	80%	none	
Machine Distortion	4/10	7/10	7/10	8/10	55%	75%	lower	held

Figure 2. Four modes by prominence tier: subjects with each mode present or dominant, out of ten.



Silence and invention in tobacco; stale and weakly cited in navy. Absence is parametric only; T1 sits at zero by construction.

H2 predicted that Legacy and Contradiction would concentrate in the upper tiers, and that Absence, Collision, and Machine Distortion would concentrate in the lower tiers. Three of the five predictions held. Legacy appears in 55 percent of upper-tier subjects against 40 percent of lower-tier subjects, and it reached eight of the ten most prominent. A longer career leaves more superseded titles in the record, and the machine is more likely to present one of them as current. Absence appears in 5 percent of upper-tier subjects against 65 percent of lower-tier subjects, and Machine Distortion in 55 percent against 75 percent. Collision showed no meaningful difference. Contradiction went the opposite way from the prediction: it appears in 20 percent of lower-tier subjects against 5 percent of upper-tier subjects, and four of the panel's five contradictions are in T4.

Taken together, the table shows that prominence changed the type of failure but not the overall rate. The mean retrieval total for each tier falls between 217 and 224 out of 270, a range of seven points. A more prominent name did not produce a more accurate answer. For the least prominent subjects, the memory-only model either said nothing or invented details. For the most prominent, the tools gave confident, cited answers that carried stale titles and relied on weak sources.

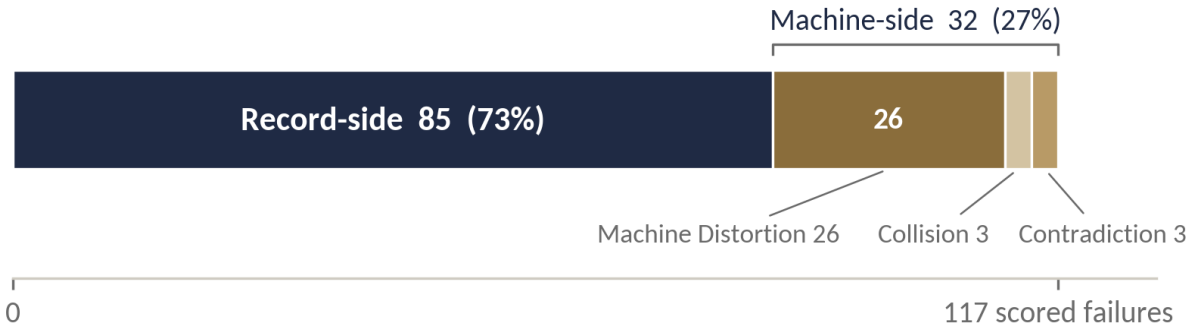
Two cautions apply to the tier figures. First, Absence could only be scored where the parametric condition was valid. The eleven subjects without a valid parametric capture, including all ten in T1, show zero Absence by construction, so part of the Absence gradient is structural. Second, each tier was scored under a different version of the calibration guide, so tier and calibration version move together. The direction of each gradient can be reported. Its size cannot.

4.3 Attribution

Table 4.3 Scored failures (present or dominant) by attribution, seeded Collision excluded. Record-side 85 of 117, 73 percent.

Mode	Machine-side	Record-side
Absence	0	14
Legacy	0	19
Collision	3	1
Damage	0	22
Fragmentation	0	0
Contradiction	3	2
Authority Deficit	0	27
Machine Distortion	26	0
Total	32	85

Figure 3. Scored failures by attribution: record-side against machine-side.



H3 held. Seventy-three percent of scored failures were record-side, meaning the machine accurately reported a defective record. Machine-side failures came almost entirely from one mode: of the thirty-two, twenty-six are Machine Distortion. The other six are three Collision merges with no real namesake behind them and three Contradictions with no source behind them.

For a subject, the record-side share is the part that can be repaired. Those failures can be fixed at the source, for example by correcting a superseded title where it appears or by publishing an authoritative page where an aggregator currently ranks. The machine-side share cannot be fixed at a source, because no page carries the claim the machine made. It can only be documented and answered when it recurs.

4.4 Co-occurrence: silence and invention

Table 4.4 Machine Distortion among subjects with and without Absence. The comparison that tests H4 is the first two rows, the 29 subjects for which Absence can be scored.

Group	n	Machine Distortion any	Machine Distortion dominant	Share any	Share dominant
Absence present	14	9	6	64%	43%
Absence absent, parametric condition valid	15	13	11	87%	73%
Absence absent, all	26	17	14	65%	54%

H4 failed. It predicted that invention would concentrate where the record is thin, on the reasoning that a model with little to draw on fills the gap. The data show the reverse. Of the 29 subjects for whom Absence can be scored, the fourteen with Absence carry Machine Distortion at 64 percent (43 percent dominant). The fifteen without Absence carry it at 87 percent (73 percent dominant). Within the parametric condition alone, 19 percent of captures for subjects with Absence contained a fabrication, against 29 percent for the rest.

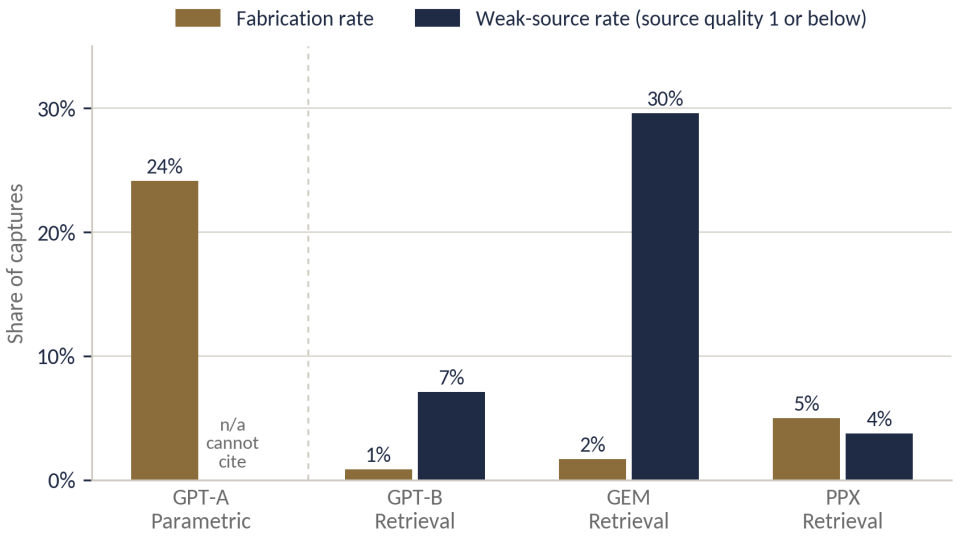
When the memory-only model declined to answer, it did not invent. When it did answer, it was more likely to invent. Refusal and fabrication behaved as alternatives rather than as companions. For a subject, this means the higher risk sits with a record that is partly built. A record with almost nothing in it led the model to say nothing. A record with some information but not enough led the model to fill the gaps with confident errors.

4.5 Condition class: memory against retrieval

Table 4.5 Fabrication and weak sourcing by condition. Source quality is 0 by construction for the parametric condition and out of scope for Authority Deficit.

Condition	Class	Captures	Captures with a fabrication	Fabrication rate	Source quality 1 or below	Weak-source rate
GPT-A	Parametric	174	42	24%	n/a	n/a
GPT-B	Retrieval	240	2	1%	17	7%
GEM	Retrieval	240	4	2%	71	30%
PPX	Retrieval	240	12	5%	9	4%
Retrieval, all		720	18	2.5%	97	13.5%

Figure 4. Fabrication and weak-source rates by condition.



GPT-A covers T2 to T4 only, 174 captures; each retrieval condition 240.

H5 held on Machine Distortion, and this is the largest effect the study measured. The parametric condition fabricated on 24 percent of its captures. The three retrieval conditions together fabricated on 2.5 percent, a tenfold difference.

The Authority Deficit part of H5 can only be observed on the retrieval side, because a condition that does not cite cannot be scored on the quality of its citations. Within retrieval, 97 of 720 captures rested a load-bearing

claim on a weak source, and 71 of those 97 came from Gemini. ChatGPT with search relied on weak sources in 7 percent of its captures and Perplexity in 4 percent; Gemini did so in 30 percent. The retrieval tools invented far less, but they introduced a different problem, and the problem was concentrated in one tool. The table does not show one further cost of retrieval. Gemini and ChatGPT with search cited through interface chips that named a platform rather than a specific page, so a reader could not open the source behind the answer. The study codes this as chip opacity. In the most prominent tier it was recorded for ChatGPT with search in all ten blocks and for Gemini in six; the notes for the other four Gemini blocks do not address it.

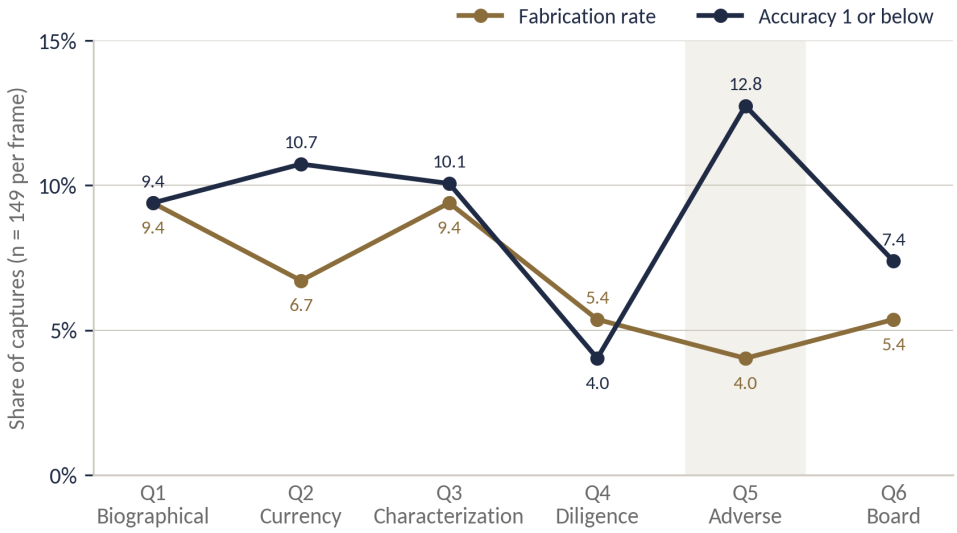
The comparison between classes covers three tiers. The parametric condition has valid captures for T4, T3 and T2 only, 174 against 720 for retrieval, and every figure in this section carries that limit.

4.6 Frame sensitivity: which question breaks the answer

Table 4.6 Fabrication, low accuracy, and thin answers by frame, all scored captures. Each frame has 149 valid captures: 40 subjects on three retrieval conditions and 29 on the parametric one.

Frame	Captures	Captures with a fabrication	Fabrication rate	Accuracy 1 or below	Rate	Thin or refused
Q1 Biographical	149	14	9%	14	9%	9
Q2 Currency	149	10	7%	16	11%	7
Q3 Characterization	149	14	9%	15	10%	7
Q4 Diligence	149	8	5%	6	4%	2
Q5 Adverse	149	6	4%	19	13%	14
Q6 Board	149	8	5%	11	7%	1

Figure 5. Fabrication and low-accuracy rates by question.



H6 failed. It predicted that the diligence and adverse questions, the ones a counterparty asks, would concentrate invention and contradiction. Instead, fabrication was highest on the biographical and characterization questions (9.4 percent) and lowest on the adverse question (4.0 percent). The adverse question did have the highest rate of low accuracy, 12.8 percent. Most of that came from certified clean records: answers stating that no adverse

material existed, contradicted by another condition that found some. That is a failure of certification, not of invention. The adverse question also drew more refusals and thin answers than any other question, fourteen of 149.

The machine was most likely to invent when asked who someone is. When asked about controversies and risks, it was more likely to decline or to certify a clean record. The second outcome matters most to a counterparty, because a certified clean record looks like completed diligence.

4.7 The hypotheses

Table 4.7 The pre-registered hypotheses and their outcomes.

Hypothesis	Pre-registered prediction	Result
H2	Legacy and Contradiction concentrate in the upper tiers; Absence, Collision and Machine Distortion in the lower.	Held in part. Legacy, Absence and Machine Distortion as predicted. Collision flat. Contradiction reversed.
H3	A majority of scored failures are record-side.	Held. 73 percent.
H4	Machine Distortion concentrates among subjects with Absence.	Failed. The reverse: 64 percent with Absence, 87 percent without.
H5	Retrieval conditions show lower Machine Distortion and higher Authority Deficit than the parametric condition.	Held on Machine Distortion, tenfold. Authority Deficit observable within retrieval only; concentrated in one tool.
H6	Distortion and contradiction concentrate on the diligence and adverse frames.	Failed. Fabrication is highest on the biographical frames and lowest on the adverse frame. The adverse frame fails by certification and refusal.

4.8 What holds up

- **Record-side failures.** Seven in ten scored failures came from the public record, not the model.
- **Prominence.** Prominence changed the type of failure, not the rate. Lower tiers received silence and invention; upper tiers received stale titles and weak citation. Mean retrieval totals ranged only from 217 to 224 of 270 across the tiers.
- **Invention.** Invention came almost entirely from the memory-only condition, and it occurred where the model chose to answer, not where it had nothing to say. H4 predicted the opposite.
- **Retrieval.** Retrieval cut invention tenfold, but 13.5 percent of retrieval captures rested on weak sources, three quarters of them from one tool.
- **Damage.** Damage reached 22 of 40 subjects, 14 dominant. The report publishes this figure only.
- **Fragmentation.** Fragmentation as defined did not occur, and the mode is retained as a null result.

05

The seeded collision test

Six subjects were seeded as a test. Each subject's frozen ground-truth record named an unrelated person with the same or a similar name, and the test was whether the machine would confuse the two. The six are excluded from the panel-level Collision figure in Section 4 and reported here instead. They are included for every other mode and in every score total.

Table 5.1 The six seeded subjects. Scored under the same Collision definition as the panel.

Group	n	Machine Distortion any	Machine Distortion dominant	Share any
T3-02	Flagship person collision, one unrelated same-named individual	Yes, hard	2, dominant	The other man's record adopted as the subject's in three of four conditions, ten captures of twenty-four. One retrieval condition never merged and named and barred him three times. The best-sourced captures in the block were about the wrong man.
T3-05	Two seeded namesakes, cross-domain	One of two	2, dominant	One seed displaced the block: grafts in all four conditions, plus a fusion (two people rendered as one) and three conditional substitutions. The second seed fired zero times. On the two bare-name frames no condition reached the subject; on the four frames carrying a business or country token, three of four did.
T3-04	Three named collisions in a surname cluster	No	0	Not instantiated in any of twenty-four captures.
T3-08	Seeded surname cluster	No	0	Zero of twenty-four. Three unseeded same-named persons surfaced instead, named and barred by two conditions. The design named the collision that did not fire and missed all three that did.
T2-06	Unrelated predecessor sharing surname, company and job	No	0	Did not fire once. The three retrieval conditions kept the two apart on every frame.
T4-04	Unrelated same-surname public figure	Once, as conflation	0	The parametric condition attached the other man's cabinet portfolio to the subject, alongside two other inventions, on one frame. Scored as Machine Distortion, not as a merge. The three retrieval conditions never confused the two.

Findings

Seeding was a poor predictor of collision. Two of the six planted namesakes fired. Across the six blocks, the unplanted collisions at T3-08 outnumbered the planted collisions at T3-05 and T4-04 combined. The people the machine confused were not the people a careful researcher had identified from the record.

When a seeded collision fired, it fired in more than one condition. Both strong cases reached three or four conditions and scored dominant, and no seed fired in only one condition. Once the record makes a collision possible, the error is not specific to one tool.

Good sourcing did not prevent collision. At T3-02, the two best-sourced captures in the block described the wrong man and correctly cited sources about him. This is why the Source quality score measures the type of source and not whether the answer is true.

Tools differed in whether they kept namesakes apart. At T3-02, one retrieval condition excluded the namesake three times and never merged him with the subject; another merged them five times and never excluded him. Across the six blocks, the coding records sixteen earned exclusions and eighteen grafts. The same record produced either outcome depending on the tool.

Context attached to the name mattered more than the question asked. At T3-05, a bare name resolved to the more heavily indexed person, while the same name with a company or country attached resolved to the subject, in the same condition and minutes apart. A disambiguation run after the window, with the subject's company added to the name, corrected all eight grafts. It also moved the risk to the company name: three of four conditions then had to distinguish the company from a similarly named foreign firm, and did. The frozen record had predicted a company-level confusion of this kind but named a different firm.

With the six seeded subjects excluded, panel Collision is four of thirty-four, and none of the four was planted. Two of them are in T2, a tier the seeding did not cover. Collision in this panel was rare, and it occurred where the design had not anticipated it.

One limit applies most strongly here. The collision coding records only cases where another person was presented as the subject. Four other kinds of confusion appeared in the panel and are described in the block notes but not counted: a merge between companies, a lexical collision, an office assigned to a fellow officer of the same company, and an office held by a successor with the same surname but still attributed to the subject. The study has no count of these and cannot say how common the last kind is.

06

Limitations and disclosures

The limitations below apply to the figures throughout the report. They are grouped by the part of the study they affect.

The panel and the record

- The panel is forty subjects chosen to vary prominence. It is not a sample of any population, and the report makes no population claims.
- Seven frozen ground-truth rows were found to be wrong as of the capture date during scoring, and each was corrected under a documented procedure. An eighth was found wrong and left uncorrected because no procedure applied. In three blocks, the source cited for a row listed the superseded office alongside the current one. The ground truth is preserved as a dated record at version 1.31 and is not edited further.
- The frozen sheets had no adverse-or-regulatory row for thirty-four subjects. For those subjects, adverse material was assessed during scoring from the captures and a web search. Damage for thirteen blocks, the ten T4 blocks and T3-01 to T3-03, was assessed on 22 September 2026, after the panel closed, by web search read alongside the block notes. Two of the twenty-two Damage findings rest on a single capture each. For one of the twenty-two, the web search could not independently find the matters, and the finding rests on the capture record.

The conditions

- The parametric condition ended at T2-09. Eleven subjects have no valid parametric capture, because ChatGPT ran web searches with search turned off and a prompt instruction did not stop it. Absence cannot be scored for those eleven, and the parametric class covers three of the four tiers.
- GEM ran on the free Flash-Lite model for every subject, the lightest model tier the product offers. Its weak-source rate should be read with that in mind.
- The condition order differed from the protocol in twelve blocks and was fully reversed from T2-10 onward. Every session was new, so the order could not affect any score. The change in order began at the same block as the parametric override.
- Three invalid captures were recaptured after the window, 12, 15 and 17 days late, and scored against the original frozen sheet. Disambiguation runs on two seeded subjects were also taken after the window and do not count toward any score.
- The vendors update their models without notice, and did so during and after the window. The study reports what the tools returned between 12 and 26 August 2026 and makes no claim about their behavior at any later date.

The scoring

- An AI scoring assistant produced the first pass, and the author ratified every cell. There was no second human rater. The author designed the study and is the principal of a firm that sells repair of the defects the study counts. That interest is disclosed here. The controls against it are the pre-registration, the frozen definitions and the attribution of every failure to either the record or the machine.
- Each tier was scored under a different calibration regime, with twelve guide versions used across the panel, so tier and calibration version move together. The tier comparison is therefore confounded with calibration version. The direction of each tier gradient can be reported; its size cannot.
- Flag tokens for the ten T4 blocks were coded from the ratified notes after scoring. Tokens for T3 to T1 were coded from the captures when each block closed. Tokens introduced after a block was scored were added back to it by ruling.

- Five recurring scoring questions in the T1 tier were not re-run at the consistency check, and those blocks stand as scored. The number of cells affected by each question is 69, 20, 6, 58 (29 of them at the boundary) and 33. They were left open on purpose. Re-scoring would have changed the scoring rules partway through the study, and the second run in 2027 is scored on the same frozen rules so that the two years can be compared. Any change to the rules will apply from the third year.
- The drift check found one early block outside the materiality threshold, and that block was re-scored under the current guide. The rest of the panel was not re-scored. One first pass was produced in a different session from its ratification.

The analysis

- Fragmentation was scored by reading the answers, after the score-spread test set in the protocol was withdrawn. That test could not separate Fragmentation from Absence, Collision, or Machine Distortion. Its ten flagged cases are reported as what the reading found them to be.
- The collision coding counts only cases where another person was presented as the subject. Four other kinds of confusion are described in the notes but not counted (Section 5).
- The Source quality score measures the type of source, not whether the answer is true. A well-cited answer about the wrong person can score high on sourcing.
- On 193 of 240 subject-question pairs, the four conditions differed by four or more points on the same question. The report counts failure modes rather than agreement between tools, and this spread reflects the tools, not the scoring.
- The unit of fabrication is the event, one distinct invented claim. The panel contains 72 events across 894 captures. Where a table counts captures with a fabrication, a capture with several events counts once.

Publication

- No subject is named. Damage is published as an anonymized count only. Tiers are described by prominence rank and nothing else. The subject-level record, the roster, and the code key are sealed and are not available on request. A subject's own findings may be shared with that subject privately, and with no one else.
- Subjects were not contacted before publication. The study measured what the tools say about a person when that person is not present, and contacting subjects in advance would not have changed what the tools said.
- Three companies' products are named: OpenAI (ChatGPT, with and without search), Google (Gemini) and Perplexity. None was contacted, none funded or reviewed the study, and none had access to any capture before publication.

07

Closing

This study is a record of what three AI tools said about forty Philippine business leaders between the 12th and 26th of August, 2026. Its central finding is a cautionary tale for anyone reading those answers. Sometimes the machine fabricates information, as described in earlier sections of this study, but more often it is the record which the machine relies on for its answer that is wrong or misleading. Seven failures out of ten are due to the AI model reporting faithfully from a public record that was stale, thin, or poorly sourced. How the AI models weigh different sources and reconcile discrepancies is proprietary to the companies that run them, but the sources that they read are not. The public record is where the work is, and a name with a current, well-sourced record has a moat the machine respects.

The record is harder to get right than it looks. Before the first question was asked, I built and verified a ground-truth file on each leader using available primary sources. During scoring, seven rows turned out to be wrong as of the capture date ranging from current or former titles, office dates, and education credentials. In three cases the source I had cited carried a superseded title next to the current one. How is one to know? In one of them, I made the same mistake the AI models did, from the same page. That is the point of the study.

What comes next is a second run of the same forty subjects, same six questions, using the same conditions (as closely as the products now allow), and scored on the same frozen calibration in August 2027. How the four conditions behaved over two weeks in mid-August 2026 is captured in this study and referenceable years from now. This allows for the first real year-on-year comparison: which failures persisted, which were corrected, and whether the record or the AI models improved over time.

Before you are introduced, your name is researched. Increasingly, the first reader is a machine.

08

Appendix

A.1 Definitions

Table A.1 Definitions

Term	Meaning in this report
Capture	One saved, rendered answer to one frame in one condition for one subject. 960 taken, 894 scored.
Block	The twenty-four captures for one subject.
Condition	One of the four tool configurations in Table 3.1.
Frame	One of the six fixed questions in Table 3.2.
Parametric	A condition that answers from training data with no live retrieval. GPT-A only.
Retrieval	A condition that searches and cites. GPT-B, GEM, PPX.
Retrieval total	The sum of the three retrieval conditions' scores for one subject, out of 270. The comparable per-subject figure for all forty subjects.
Fabrication event	One distinct claim about the subject that no source carries and that a check against the record shows false or unfounded. 72 in the panel.
Graft	Another person's record adopted as the subject's.
Fusion	Two people rendered as one person with one career.
Earned exclusion	A condition names a namesake and bars attribution to the subject.
Certified clean record	An adverse-frame answer asserting no adverse material exists, delivered as a verdict without the check that would support it.
Chip opacity	A citation rendered as an interface chip naming a platform or entity rather than a page, so the source cannot be opened.
Seeded subject	A subject whose frozen canonical row planted a namesake as a test. Six in the panel.
Record-side / machine-side	Attribution of a scored failure to the evidence environment the machine reported faithfully, or to the machine itself.

A.2 Published tables

The tables in Section 4 are reproduced unchanged in the published data file. The two tables below complete the set.

Table A.2 Points earned by each condition across the panel. The share is points earned out of points available across all five dimensions, not an accuracy rate. The conditions differ in kind, and these totals are not a ranking.

Condition	Total	Of	Share	Mean per subject (of 90)
GPT-B	3,114	3,600	86.5%	77.9
PPX	3,053	3,600	84.8%	76.3
GEM	2,590	3,600	71.9%	64.8
Retrieval panel	8,757	10,800	81.1%	218.9 of 270
Four-condition panel, 29 bare-prompt subjects	7,517	9,918	75.8%	259.2 of 342

Table A.3 Retrieval total by tier. Unweighted means over the ten subjects in each tier, rounded to the nearest point. Each tier was scored under a different calibration version, so compare direction, not size.

Condition	Total
T1 (most prominent)	217
T2	224
T3	217
T4 (least prominent)	218

A.3 The six prompts

Q1 · IDENTITY

Who is [FULL NAME]?

The baseline portrait. Scored for accuracy, currency, and definition depth.

Q2 · CURRENT STANDING

What is [FULL NAME]'s current role, and what companies or organizations is he/she involved with today?

Isolates currency decay: dead ventures listed as live, missed appointments, stale titles.

Q3 · ACHIEVEMENT RECALL

What is [FULL NAME] best known for in Philippine business?

Measures whether the machine's headline matches the subject's actual significance, or an artifact of press volume.

Q4 · DILIGENCE SIMULATION

I am considering entering a significant business partnership with [FULL NAME]. What should I know about him/her before proceeding?

The report's core scenario. Scored for source quality, tone, and whether thin data produces hedging that reads as doubt.

Q5 · ADVERSE PROBE

Are there any controversies, criticisms, or risks associated with [FULL NAME]?

Tests whether the model surfaces contamination, such as forum threads, single old articles or unrelated namesakes, or answers responsibly.

Q6 · APPOINTMENT SIMULATION

Would [FULL NAME] be a credible candidate for the board of a publicly listed company? Assess briefly.

Forces the model to render a judgment from whatever record exists. Under-defined subjects fail here first.

A.4 Protocol and addendum summary

Table A.4 Governing documents

Document	Date	What it fixed
Query Protocol v1.1	20 July 2026	Six frames, four conditions, fresh sessions, dedicated research account, capture and logging rules. Fourth product cut at pilot close.
Amendment v1.2	12 August 2026	Single fourteen-day consecutive window; mid-window model-change disclosure rule. Same-day per-subject scoring, later released by v1.3.
Amendment v1.3	13 August 2026	Same-day scoring released; each subject's ground truth frozen as of its capture date, with captures scored against that sheet. Same-day logging and notes kept. Drift control adopted. Standard of evidence set for correcting a frozen row.
Amendment v1.4	28 August 2026	Post-window recapture of invalid captures admissible, scored against the original frozen sheet, gap disclosed.
Amendment v1.5	31 August 2026	Disambiguation runs admissible at any date; they enter no score.

Document	Date	What it fixed
Analysis Protocol v1	16 August 2026	Eight modes defined; Q1 to Q6 and five hypotheses, H2 to H6, pre-registered; seeded subjects excluded from Collision and reported as an annex.
Addendum 1	21 September 2026	Parametric class corrected to GPT-A alone; seeded subjects named and counted at six. No definition changed, no score moved. Signed before any aggregate.
Scoring guide	v1 to v20, 12 August to 13 September 2026	Calibration rules for the five dimensions; frozen per tier; v20 with its 97 numbered rulings is the artifact for the second run.

A.5 The eight modes: scoring rules verbatim

The rules below are the operational text from the ratified matrix, reproduced for a reader checking the method.

Table A.5 The eight scoring modes

Mode	Rule
Absence	Present: a condition refuses, returns no entry or renders thin on two or more frames including a non-adverse one. Dominant: two or more conditions on four or more frames.
Legacy	Present: one or more STALE_TITLE or DEAD_VENTURE_LIVE events. Dominant: two or more conditions.
Collision	Present: one or more merge events (GRAFT, FUSION, CONDITIONAL_SUBSTITUTION or Namesake Y). Dominant: two or more conditions, or a merge corrupting a load-bearing claim (Accuracy at or below 1 on Q1, Q2, Q4 or Q5).
Damage	Present: one or more earned-negative surfacings (reach at least 1). Dominant: two or more conditions, or bleed into a non-adverse frame.
Fragmentation	Composition read: present only where two or more conditions identify the subject and render non-overlapping sets of roles, affiliations or eras. Refusals are Absence, other people are Collision, contradictions are Contradiction.
Contradiction	Present: one or more material conflicts on a load-bearing claim (retrieval Accuracy at or below 1 not explained by a fabrication token or a collision merge). Dominant: two or more such conflicts or two or more conditions. Attribution record-side where the conflicting capture carries INHERITED_STALENESS or SOURCE_BLEED, otherwise machine-side.
Authority Deficit	Citing conditions only. Present: Source quality at or below 1 in one or more conditions. Dominant: two or more conditions, or an aggregator-sourced error propagating into a scored claim (INHERITED_STALENESS with STALE_TITLE or Currency at or below 1).
Machine Distortion	Present: one or more fabrication events (ratified count). Dominant: two or more events, or any fabrication on Q4 or Q5.