The Apophenia Gradient

Do language models tell genuine semantic connection from coincidence — and does it improve with capability? A signal-detection study

Alejandro Toledo Martínez

A companion, AI-facing study to the endolinguistic work in ../specular-cognation/. That project needed a judge of whether two words are *genuinely* connected in sense, and found that language models will, under the wrong prompt, affirm almost any connection. This paper isolates that tendency — a machine analogue of apophenia — measures it with signal-detection tools across a capability range, and reads the result as a small, concrete facet of the hallucination problem. It supports the endolinguistic papers (it says which judges can be trusted, and why) and stands on its own as a measurement.

Abstract

Apophenia is the perception of connection where there is none. A language model asked whether two words are related in meaning is a natural apophenia machine: prompted for a yes/no verdict, a mid-size model (gemma2) calls *every* pairing natural, narrating a plausible bridge for each. We ask whether this is intrinsic or a function of capability, using a clean 43-item battery with a 2×2 design (shared code × genuine/forced link) and a graded 0–100 rubric that neutralises the yes-to-everything bias. Across a panel of eight judges the discrimination of genuine from forced links (AUC) rises monotonically with capability — qwen2.5-3b 0.58 (chance) → gemma2-9b 0.84 → Claude 0.91–0.96 — and the criterion tightens in step: capable judges score spurious links far lower (forced-link mean 17 → 5) while raising genuine ones. The gradient is not an artefact of the model-derived item key: re-scored against a human-consensus key (twelve naive raters, Cronbach $\alpha = 0.93$), discrimination still rises with capability, and the human raters themselves sit mid-gradient (AUC 0.82). Two findings sharpen the picture. First, a small model is not apophenic but *noisy*: qwen scores everything 18, discriminating nothing. Genuine apophenia — high confidence in spurious links — is a middle-capability, binary-prompt phenomenon, and graded scoring plus capability dissolves it. Second, deliberation does not help: a snap verdict and a careful one give the same AUC (0.95 → 0.96); apophenia is not cured by thinking longer but by capacity. As a by-product the battery confirms an endolinguistic negative: at equal link quality a shared consonantal code affords nothing (in-code = decoy). The corollary is a caution: the same capability that sharpens the judge also lets it *generate* more convincing spurious links — the danger is not that strong models see connections everywhere, but that the connections they manufacture are hard to refute.

Resumen

La apofenia es percibir conexión donde no la hay. Un modelo de lenguaje al que se le pregunta si dos palabras se relacionan en sentido es una máquina de apofenia natural: pedido un veredicto sí/no, un modelo mediano (gemma2) llama *natural* a *todo* par, narrando un puente plausible para cada uno. Preguntamos si esto es intrínseco o función de la capacidad, con una batería limpia de 43 ítems de diseño 2×2 (código compartido × enlace genuino/forzado) y una rúbrica graduada 0–100 que

neutraliza el sesgo del sí-a-todo. En un panel de ocho jueces la discriminación de enlaces genuinos vs forzados (AUC) sube monotónicamente con la capacidad — qwen2.5-3b 0,58 (azar) → gemma2-9b 0,84 → Claude 0,91–0,96— y el criterio se endurece a la par: los jueces capaces puntúan los enlaces espurios mucho más bajo (media forzado 17 → 5) mientras suben los genuinos. El gradiente no es un artefacto de la clave derivada de un modelo: re-puntuado contra una clave de consenso humano (doce evaluadores naive, Cronbach $\alpha = 0{,}93$), la discriminación sigue subiendo con la capacidad, y los propios evaluadores humanos caen a mitad del gradiente (AUC 0,82). Dos matices afinan la imagen. Primero, un modelo pequeño no es apofénico sino *ruidoso*: qwen puntúa todo 18, sin discriminar nada. La apofenia genuina —alta confianza en enlaces espurios— es un fenómeno de capacidad media y prompt binario, y la puntuación graduada más la capacidad lo disuelven. Segundo, deliberar no ayuda: un veredicto instantáneo y uno cuidadoso dan el mismo AUC (0,95 → 0,96); la apofenia no se cura pensando más sino teniendo capacidad. Como subproducto la batería confirma un negativo endolingüístico: a igual calidad de enlace, un código consonántico compartido no aporta nada (en-código = señuelo). El corolario es una cautela: la misma capacidad que afila al juez le permite *generar* enlaces espurios más convincentes — el peligro no es que los modelos fuertes vean conexiones por todas partes, sino que las que fabrican son difíciles de refutar.

Keywords / Palabras clave

apophenia, hallucination, signal detection, semantic connectivity, generator–judge gap, model capability, LLM evaluation.

1. The machine that says yes

A language model is, among other things, a device for finding connections. Ask one whether two arbitrary words — *soltar* (to let go) and *mesa* (table) — are related, and it will oblige: *"you release your grip on the table."* Ask about *soltar* and *dog*: *"you let the dog off the leash."* About *soltar* and *candle*: *"a candle releases wax as it burns."* Each bridge is fluent and each is, on reflection, forced. This is the machine analogue of apophenia — the perception of connection where there is none — and it is not a rare failure mode but the default when the model is asked for a binary verdict. A mid-size open model (gemma2) never once answered *not related* on such pairs.

The tendency matters beyond a curiosity. Apophenia is one root of a certain kind of hallucination: the model "sees" a relation that is not there and states it with confidence, wrapped in a plausible rationale. And it bears directly on a companion project — the endolinguistic "code game" in [`../specular-cognation/`](../specular-cognation/) — which needed a *judge* of whether two words are genuinely connected, and could not use a judge that says yes to everything. So we ask the question sharply: do language models discriminate genuine semantic connection from coincidence, and does that ability scale with capability?

2. Hypothesis — a generator–judge gap that scales

A model that produces a connection and a model that evaluates one draw on different competences. Generating a plausible bridge is cheap; recognising that the bridge you just generated is nonetheless forced is a second, metacognitive act. We call the distance between them the generator–judge gap, and hypothesise that it *widens with capability*: weak models generate freely and judge poorly; strong models can hold *"I can produce this link and it is still forced."*

In signal-detection terms this predicts two movements as capability rises: higher sensitivity (better

separation of genuine from spurious links, d' / AUC) and a stricter criterion (less willingness to score a spurious link as strong). The alternative worth taking seriously is that larger models merely generate *more convincing* bridges — apophenia that is harder to catch, not less of it. The battery is built to tell these apart.

3. Design — a clean battery, graded scoring, a panel

The battery. 43 items in English, each a (seed, candidate) pair with common words and common senses, arranged in a 2×2: whether the candidate shares the seed's folded consonantal code (*in-code* vs *decoy*), crossed with whether the conceptual link is *genuine* (metonymy/metaphor: cause, instrument, part, symbol) or *forced* (only co-occurs, or far-fetched). The code is computed and verified for every item. The design lets one measurement do double duty: the genuine/forced axis reads apophenia; the code axis tests, as a by-product, whether a shared code affords connection at all. Adversarial *decoy-genuine* items (a real link whose code does *not* match — *fire→smoke*, *sun→star*) guard against a judge crediting the code rather than the sense.

Graded scoring. The binary question saturates — it invites the yes-to-everything reflex. Each judge instead rates 0–100 how genuine the metonymic/metaphoric relation is, under a strict rubric calibrated with worked examples (*cut act* ≈ *70; cut cat* ≈ *8*), told that most random pairings must fall below 25.

The panel. Eight judges spanning two orders of capability: Ollama-served qwen2.5-3b and gemma2-9b; and Claude Haiku 4.5, Sonnet 5, Fable 5, and Opus 4.8, the last two in a *snap* (immediate) and a *careful* (deliberate) mode. Measures: per judge, AUC for genuine-vs-forced separation (sensitivity), the mean score given to *forced* links (a proxy for criterion — how strong the judge rates the spurious), and the in-code-vs-decoy comparison at equal link quality (the code effect). The item key here is one strong model's labelling; eight independent judges reproduce the ordering, and a human panel (§4) confirms it against an independent standard.

4. Results — the gradient

Discrimination rises monotonically with capability, and the criterion tightens with it.

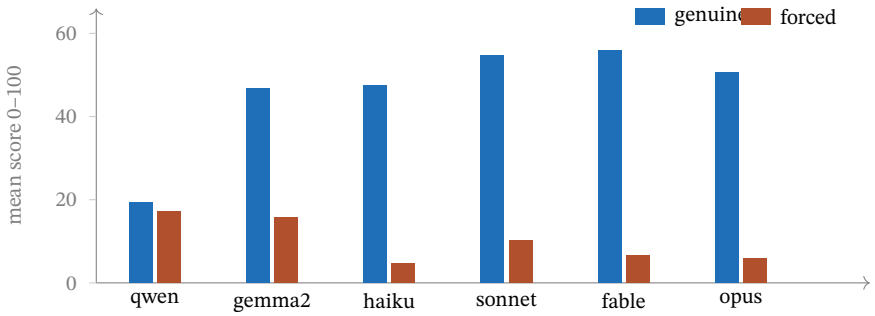

Figure 1: Mean score given to genuine vs forced links, by judge (ordered by capability). The *gap* is sensitivity; the height of the *forced* bar is the criterion (how strong the judge rates spurious links). Capable judges both open the gap and press the forced bar down. *Puntaje medio a enlaces genuinos vs forzados, por juez (ordenados por capacidad). La* brecha *es la sensibilidad; la altura de la barra* forzado *es el criterio. Los jueces capaces abren la brecha y hunden la barra de forzados.*

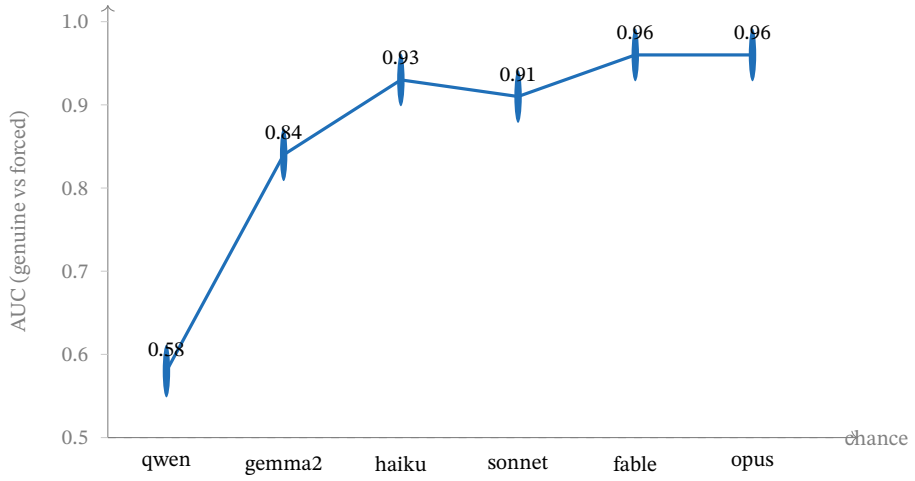

Figure 2: Discrimination (AUC, genuine vs forced) rises with capability and plateaus near 0.96. qwen sits at chance; the jump from qwen to gemma2 is the largest. *La discriminación (AUC) sube con la capacidad y se estabiliza cerca de 0,96. qwen está en el azar; el salto qwen→gemma2 es el mayor.*

judge	params	AUC	genuine	forced	in-code (forced)	decoy (forced)
qwen2.5	3B	0.58	19.3	17.2	16.1	20.0
gemma2	9B	0.84	46.7	15.7	15.7	15.7
Haiku	4.5	0.93	47.4	4.7	5.4	2.9
Sonnet	5	0.91	54.6	10.2	9.7	11.4
Fable (snap→careful)	5	0.95→0.96	55.8	6.5	6.5	6.6
Opus (snap→careful)	4.8	0.96	50.6	5.9	6.2	5.0

Three readings. (1) Sensitivity scales. AUC climbs from chance (qwen 0.58) to 0.96, the largest step being the first — small-to-mid. (2) The criterion tightens with it. The mean score awarded to *forced* links falls from 17 (qwen) to 5 (Claude): capable judges do not merely separate genuine from spurious, they actively refuse the spurious. (3) Deliberation is not the lever. Fable's snap and careful modes give 0.95 and 0.96 — the same, within noise; careful mode sharpens the scores (genuine up, forced down) without moving the discrimination. For a capable judge the intuitive verdict is already as good as the deliberate one.

A nuance corrects the original "gemma2 is pure apophenia" impression. That impression came from *binary* judging of *garbage* senses; on clean items with graded scoring gemma2 discriminates at 0.84. And the weakest model is not apophenic at all: qwen scores everything 18 (genuine 19.3, forced 17.2) — it is *noisy*, not credulous; it fails to discriminate rather than over-affirming. Genuine apophenia — confident endorsement of spurious links — lives in the middle, and is dissolved by graded scoring plus capability.

Finally, the by-product for the endolinguistic project: at equal link quality (*forced*), in-code and decoy candidates score the same (gemma2 15.7 = 15.7; the others within a couple of points, in either direction). A shared consonantal code affords no extra semantic connection. The adversarial decoy-genuine items are scored *high* by every capable judge, confirming they credit the sense, not the code.

A human key. The obvious worry is circularity: the genuine/forced key is itself a strong model's labelling, so the top models might score well by agreeing with their own kind. To check, we ran the 43-item battery past naive human raters — shown only the word pairs and their glosses, blind to code, condition and label — and, filtering by agreement with the consensus (leave-one-out correlation), kept twelve raters whose ratings cohere at Cronbach $\alpha = 0.93$. Re-scoring every judge's AUC

against the *human* consensus key rather than the model key, the gradient holds: qwen 0.59, gemma2 0.80, Claude 0.89–0.93, and the rank correlation of a judge's scores with the human mean rises with capability (Spearman ρ 0.14 $\rightarrow$ 0.81). The human raters, scored leave-one-out against one another, sit at AUC 0.82 — mid-gradient, between gemma2 and the strongest Claude judges. The capability gradient is therefore not an artefact of a model-derived key: it survives an independent human standard, and humans occupy the very axis they define. (This is a pilot at twelve raters; more will sharpen the absolute values, not the ordering.)

5. Discussion

Apophenia is a function of capacity, not effort. The gradient and the snap/careful equivalence together say something specific: the ability to tell a genuine connection from a manufactured one is a property of *how capable* the model is, not of *how hard it tries*. This is unusual — for many tasks deliberation helps — and it suggests that the discrimination is not a slow, checkable computation but a fast, representational one: either the model's semantics are rich enough to place the spurious link far from the genuine, or they are not, and thinking longer does not enrich them.

The gap the study was built to see. The generator–judge gap does widen with capability, but not in the shape "big models see connections everywhere." The opposite: big models are the *strictest* judges here. The danger the alternative hypothesis named survives anyway, and moves: a capable model asked to *generate* a bridge for a spurious pair will produce a more convincing one than a weak model — the §1 examples are Opus-quality rationalisations of nonsense. So the risk is not credulous judging but persuasive generation: the connections a strong model manufactures are harder for a *human* to refute, even as the model itself, asked to judge, would score them low. The practical lesson is to make models judge, not only generate — and to keep the judge separate from the generator.

What this contributes to the endolinguistic papers. It licenses the use of capable models as judges of connection (their apophenia is lowest and their criterion strictest), and it explains why the panel gradient reported in the companion paper is trustworthy rather than an artefact of one model's taste: eight judges across two orders of capability agree on the ordering, and a twelve-rater human panel ($\alpha = 0.93$) reproduces it against an independent standard. It also re-states, from the AI side, the endolinguistic negative — a shared code affords no meaning — which there is measured on words and here on a controlled battery.

> A caution the field should keep naming. In these data the most capable judges are both the most sensitive and the strictest — apophenia is not their failure mode. But the same capability is what makes their *generated* connections convincing. An evaluator that only ever reads model output, never cross-examines it, will be most misled precisely by the strongest models. Capability improves the judge and the deceiver at once.

6. Scope and limits

The item key is one strong model's labelling of genuine/forced; eight independent judges (including the much weaker qwen and gemma2) reproduce the *ordering*, and — the pre-registered check, now done (§4) — a human panel does too: re-scored against a twelve-rater human-consensus key (Cronbach $\alpha = 0.93$), the gradient holds and the raters themselves land mid-axis, so the ordering is not an artefact of a model-derived key. This human panel is a pilot; more raters will tighten the absolute AUCs, not the trend, which is why we report the gradient rather than a single number. The battery is 43 items in one language; a larger, multilingual battery would test whether the gradient's shape is language-invariant. "Semantic-connectivity judgement" is a proxy for apophenia, and apophenia is one facet of hallucination, not the whole of it. The Claude judges are non-deterministic (temperature

0.1); the Ollama judges are reproducible. The generator-side claim (§5, that capable models generate more convincing spurious links) is stated from examples, not measured here — quantifying it, ideally with humans rating the persuasiveness of model-generated bridges, is the natural follow-up.

The claim we defend is narrow and, we think, firm: discrimination of genuine from spurious semantic connection is a measurable quantity, it scales with model capability and not with deliberation, and on this axis capable language models are strict rather than credulous judges — while remaining fluent manufacturers of the very connections they would, as judges, reject.

Versión en español

1. La máquina que dice sí

Un modelo de lenguaje es, entre otras cosas, un aparato para hallar conexiones. Pregúntale si dos palabras cualesquiera —*soltar* y *mesa*— se relacionan, y complace: *"sueltas tu agarre de la mesa."* Sobre *soltar* y *perro*: *"soltaste al perro de la correa."* Sobre *soltar* y *vela*: *"una vela suelta cera al arder."* Cada puente es fluido y cada uno, pensándolo, forzado. Este es el análogo-máquina de la apofenia —percibir conexión donde no la hay— y no es un fallo raro sino lo que ocurre por defecto cuando se le pide un veredicto binario. Un modelo abierto mediano (gemma2) no respondió *no relacionadas* ni una vez.

La tendencia importa más allá de la curiosidad. La apofenia es una raíz de cierta alucinación: el modelo "ve" una relación que no está y la afirma con confianza, envuelta en una racionalización plausible. Y toca de lleno un proyecto hermano —el "juego del código" endolingüístico en [`../specular-cognation/`](../specular-cognation/)— que necesitaba un *juez* de si dos palabras están genuinamente conectadas, y no podía usar uno que dice sí a todo. La pregunta, afilada: ¿distinguen los modelos la conexión semántica genuina de la coincidencia, y esa capacidad escala con la capacidad del modelo?

2. Hipótesis — una brecha generador–juez que escala

Generar una conexión y evaluarla usan competencias distintas. Generar un puente plausible es barato; reconocer que el puente que acabas de generar es aun así forzado es un segundo acto, metacognitivo. Llamamos a esa distancia la brecha generador–juez, e hipotetizamos que *se ensancha con la capacidad*: los modelos débiles generan libremente y juzgan mal; los fuertes sostienen *"puedo producir este enlace y sigue siendo forzado."*

En términos de detección de señal esto predice dos movimientos al subir la capacidad: mayor sensibilidad (mejor separación genuino/espurio, d'/AUC) y un criterio más estricto (menos disposición a puntuar alto un enlace espurio). La alternativa a tomar en serio es que los modelos grandes solo generan puentes *más convincentes* —apofenia más difícil de atrapar, no menos—. La batería se construye para distinguirlas.

3. Diseño — batería limpia, puntuación graduada, panel

La batería. 43 ítems en inglés, cada uno un par (semilla, candidata) con palabras y sentidos comunes, en un 2×2: si la candidata comparte el código consonántico plegado de la semilla (*en-código* vs *señuelo*), cruzado con si el enlace conceptual es *genuino* (metonimia/metáfora) o *forzado* (solo coocurre, o rebuscado). El código se computa y verifica en cada ítem. El diseño hace que una medida sirva doble: el eje genuino/forzado lee la apofenia; el eje del código prueba, de paso, si compartir código aforda conexión. Ítems adversariales *señuelo-genuino* (enlace real cuyo código *no* coincide — *fire→smoke*, *sun→star*) evitan que el juez premie el código en vez del sentido.

Puntuación graduada. La pregunta binaria satura —invita al reflejo del sí-a-todo—. Cada juez puntúa 0–100 cuán genuina es la relación, con una rúbrica estricta calibrada con ejemplos (*cut act* ≈ *70; cut cat* ≈ *8*), advertido de que la mayoría de pares al azar deben caer bajo 25.

El panel. Ocho jueces en dos órdenes de capacidad: qwen2.5-3b y gemma2-9b vía Ollama; y Claude Haiku 4.5, Sonnet 5, Fable 5 y Opus 4.8, los dos últimos en modo *snap* (inmediato) y *careful* (deliberado). Medidas: por juez, AUC de separación genuino-vs-forzado (sensibilidad), la media dada a los enlaces *forzados* (proxy del criterio) y la comparación en-código-vs-señuelo a igual calidad (efecto-código). La clave es el etiquetado de un modelo fuerte; ocho jueces independientes reproducen el orden, y un panel humano (§4) lo confirma contra un estándar independiente.

4. Resultados — el gradiente

La discriminación sube monotónicamente con la capacidad, y el criterio se endurece con ella (Fig. 1, Fig. 2).

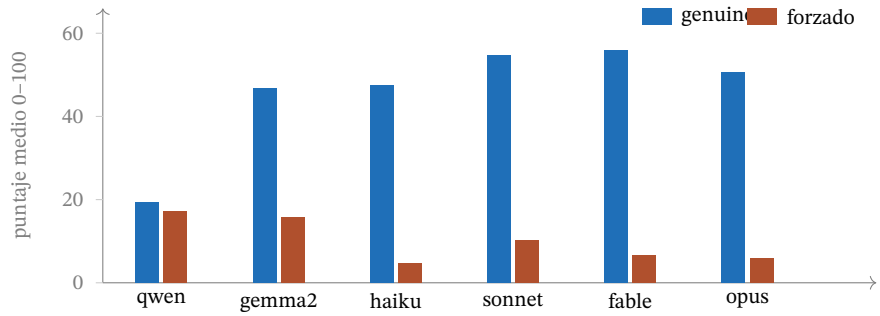

Figure 3: Puntaje medio a enlaces genuinos vs forzados, por juez (ordenados por capacidad). La *brecha* es la sensibilidad; la altura de la barra *forzado* es el criterio (cuán alto puntúa el juez lo espurio). Los jueces capaces abren la brecha y hunden la barra de forzados.

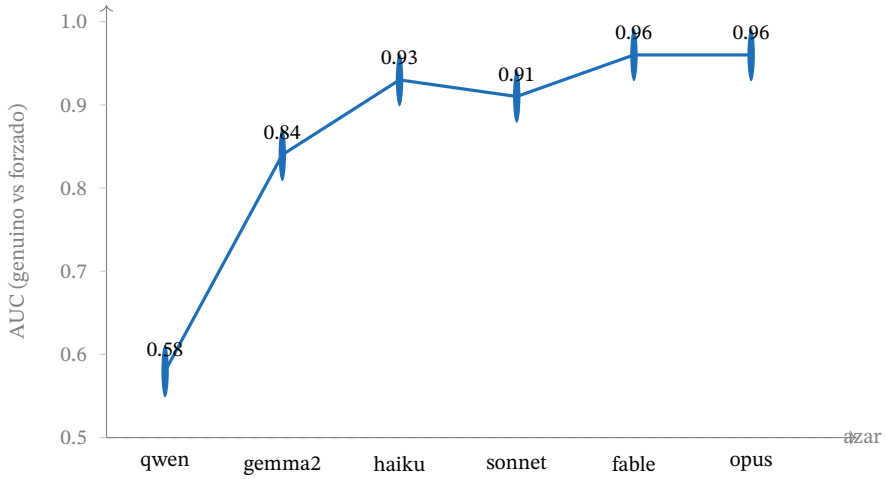

Figure 4: La discriminación (AUC, genuino vs forzado) sube con la capacidad y se estabiliza cerca de 0,96. qwen está en el azar; el salto qwen→gemma2 es el mayor.

juez	params	AUC	genuino	forzado	en-código (forz.)	señuelo (forz.)
qwen2.5	3B	0,58	19,3	17,2	16,1	20,0
gemma2	9B	0,84	46,7	15,7	15,7	15,7
Haiku	4.5	0,93	47,4	4,7	5,4	2,9
Sonnet	5	0,91	54,6	10,2	9,7	11,4

juez	params	AUC	genuino	forzado	en-código (forz.)	señuelo (forz.)
Fable (snap→careful)	5	0,95→0,96	55,8	6,5	6,5	6,6
Opus (snap→careful)	4.8	0,96	50,6	5,9	6,2	5,0

Tres lecturas. (1) La sensibilidad escala. El AUC sube del azar (qwen 0,58) a 0,96, siendo el mayor salto el primero. (2) El criterio se endurece a la par. La media dada a los enlaces *forzados* cae de 17 (qwen) a 5 (Claude): los jueces capaces no solo separan genuino de espurio, sino que rechazan activamente lo espurio. (3) Deliberar no es la palanca. Los modos snap y careful de Fable dan 0,95 y 0,96 —iguales, dentro del ruido—; el modo careful afila los scores (genuino arriba, forzado abajo) sin mover la discriminación.

Un matiz corrige la impresión original de "gemma2 = apofenia pura". Esa impresión venía de juzgar *en binario* sentidos *basura*; en ítems limpios con puntuación graduada gemma2 discrimina a 0,84. Y el modelo más débil no es apofénico en absoluto: qwen puntúa todo 18 (genuino 19,3, forzado 17,2) — es *ruidoso*, no crédulo; falla en discriminar en vez de sobre-afirmar. La apofenia genuina —endoso confiado de enlaces espurios— vive en el medio, y la puntuación graduada más la capacidad la disuelven.

Y el subproducto para el proyecto endolingüístico: a igual calidad de enlace (*forzado*), en-código y señuelo puntúan igual (gemma2 15,7 = 15,7; el resto a un par de puntos, en cualquier dirección). Un código consonántico compartido no aforda conexión semántica extra. Los ítems adversariales señuelo-genuino los puntúa *alto* cada juez capaz, confirmando que premian el sentido, no el código.

Una clave humana. La preocupación obvia es la circularidad: la clave genuino/forzado es el etiquetado de un modelo fuerte, así que los modelos top podrían puntuar bien por coincidir con los de su especie. Para comprobarlo, pasamos los 43 ítems por evaluadores humanos naive —viendo solo los pares y sus glosas, ciegos al código, la condición y la etiqueta— y, filtrando por acuerdo con el consenso (correlación leave-one-out), quedaron doce raters que cohesionan a Cronbach $\alpha = 0{,}93$. Repuntuando el AUC de cada juez contra la clave de consenso *humano* en vez de la del modelo, el gradiente aguanta: qwen 0,59, gemma2 0,80, Claude 0,89–0,93, y la correlación de rango de los puntajes de un juez con la media humana sube con la capacidad (Spearman ρ 0,14 $\rightarrow$ 0,81). Los evaluadores humanos, puntuados leave-one-out entre sí, caen en AUC 0,82 —a mitad del gradiente, entre gemma2 y los Claude más fuertes—. El gradiente de capacidad no es, pues, un artefacto de una clave derivada de un modelo: sobrevive a un estándar humano independiente, y los humanos ocupan el mismo eje que definen. (Es un piloto de doce raters; más afinarán los valores absolutos, no el orden.)

5. Discusión

La apofenia es función de la capacidad, no del esfuerzo. El gradiente y la equivalencia snap/careful juntos dicen algo específico: distinguir una conexión genuina de una fabricada es propiedad de *cuán capaz* es el modelo, no de *cuánto lo intenta*. Es inusual —en muchas tareas deliberar ayuda— y sugiere que la discriminación no es un cómputo lento y verificable sino uno rápido y representacional: o la semántica del modelo es bastante rica para poner el enlace espurio lejos del genuino, o no lo es, y pensar más no la enriquece.

La brecha que el estudio quería ver. La brecha generador–juez sí se ensancha con la capacidad, pero no con la forma "los modelos grandes ven conexiones por todas partes". Al revés: aquí los grandes son los jueces *más estrictos*. El peligro que nombraba la hipótesis alternativa sobrevive de todos modos, y se desplaza: un modelo capaz al que se le pide *generar* un puente para un par espurio producirá uno

más convincente que un modelo débil —los ejemplos de §1 son racionalizaciones de nivel-Opus de un sinsentido—. Así que el riesgo no es el juicio crédulo sino la generación persuasiva: las conexiones que un modelo fuerte fabrica son más difíciles de refutar para un *humano*, aun cuando el modelo mismo, al juzgarlas, las puntuaría bajo. La lección práctica: hacer que los modelos *juzguen*, no solo generen — y mantener el juez separado del generador.

Qué aporta a los papers endolingüísticos. Licencia usar modelos capaces como jueces de conexión (su apofenia es la más baja y su criterio el más estricto), y explica por qué el gradiente del panel del paper hermano es fiable y no un artefacto del gusto de un modelo: ocho jueces en dos órdenes de capacidad coinciden en el orden, y un panel humano de doce evaluadores ($\alpha = 0{,}93$) lo reproduce contra un estándar independiente. Y re-enuncia, desde el lado de la IA, el negativo endolingüístico —un código compartido no aforda significado— allí medido sobre palabras y aquí sobre una batería controlada.

Una cautela que el campo debe seguir nombrando. En estos datos los jueces más capaces son a la vez los más sensibles y los más estrictos —la apofenia no es su modo de fallo—. Pero la misma capacidad es la que hace convincentes sus conexiones *generadas*. Un evaluador que solo lee la salida del modelo, sin interrogarla, será más engañado precisamente por los modelos más fuertes. La capacidad mejora al juez y al embaucador a la vez.

6. Alcance y límites

La clave de los ítems es el etiquetado de un modelo fuerte; ocho jueces independientes (incluidos los mucho más débiles qwen y gemma2) reproducen el *orden*, y —la comprobación pre-registrada, ya hecha (§4)— un panel humano también: re-puntuado contra una clave de consenso humano de doce evaluadores (Cronbach $\alpha = 0{,}93$), el gradiente aguanta y los evaluadores caen a mitad del eje, así que el orden no es un artefacto de una clave derivada de un modelo. Este panel es un piloto; más raters afinarán los AUC absolutos, no la tendencia, por lo que reportamos el gradiente y no un solo número. La batería es de 43 ítems en una lengua; una mayor y multilingüe probaría si la forma del gradiente es invariante a la lengua. "Juicio de conectividad semántica" es un proxy de la apofenia, y la apofenia es una faceta de la alucinación, no toda ella. Los jueces Claude no son deterministas (temperatura 0,1); los de Ollama sí. La afirmación del lado-generador (§5) se enuncia desde ejemplos, no se mide aquí — cuantificarla, idealmente con humanos puntuando lo persuasivo de los puentes generados, es el seguimiento natural.

La afirmación que defendemos es estrecha y, creemos, firme: la discriminación de conexión semántica genuina vs espuria es una magnitud medible, escala con la capacidad del modelo y no con la deliberación, y en este eje los modelos capaces son jueces estrictos y no crédulos — aun siendo fabricantes fluidos de las mismas conexiones que, como jueces, rechazarían.

Appendix — reproducibility / reproducibilidad

Code and data under `docs/studies/llm-apophenia/`; the reduction and clean-item pipeline are shared with `../specular-cognation/` and `../transformations/`.

- `src/build_battery.py` → `data/battery.db` + `data/battery.md`: the 43-item battery, 2×2 (code-match × genuine/forced), consonantal code computed and verified via `../transformations/src/localcode.py`; adversarial decoy-genuine items included. `src/build_rating_page.py` builds a blind rating instrument (Netlify Forms / self-contained HTML) and `src/aggregate_panel.py` ingests the returned ratings, quality-filters by leave-one-out agreement, and re-scores every judge against the human-consensus key (tables `panel_human`, `panel_gradient`); the twelve-rater pilot ($\alpha = 0.93$) is reported in §4.
- `src/judge_battery.py [model]`: scores every item 0–100 via an Ollama model, writes column `score_<model>`, and computes AUC (genuine vs forced), the code effect (in-code vs decoy at equal quality), and the adversarial check. Claude judges (Haiku, Sonnet, Fable, Opus, snap/careful) were run as agents and their scores written to the same table.
- `data/battery.db` — table `battery` (one `score_*` column per judge) and table `gradient` (judge, params, auc, mean_genuine, mean_forced, in_forced, decoy_forced).

Judges: Ollama `qwen2.5:3b`, `gemma2:9b`; Claude Haiku 4.5, Sonnet 5, Fable 5, Opus 4.8. Ollama runs are deterministic; Claude runs use temperature 0.1. The battery draws its material from the endolinguistic code game (`../specular-cognation/`); the human key is the shared open task.

Código y datos en `docs/studies/llm-apophenia/`. `build_battery.py` construye la batería 2×2 (código verificado, ítems adversariales, `human_key` en blanco); `judge_battery.py` puntúa vía Ollama y calcula AUC + efecto-código + adversarial; los jueces Claude corrieron como agentes. `data/battery.db` = tabla `battery` (una columna `score_` por juez) + tabla `gradient`. Determinista en Ollama; temperatura 0,1 en Claude.