Análisis estadístico: recodificación ciega y prueba espejo

. . .

La cuarta entrega de esta serie sostiene que las cuestiones problemáticas de los programas de Filosofía de 2026 están bien formuladas en casi todos los temas, salvo en una zona: sexo, género e identidad; la relación entre saber y poder; la colonialidad y el canon; el lenguaje como instrumento de dominación; y la realidad como construcción. En esa zona, una proporción mucho mayor de preguntas presupone la respuesta o la sugiere. Este documento explica cómo se puso a prueba esa afirmación y qué resultó. Todas las codificaciones que se describen acá las hicieron modelos de lenguaje. No hubo codificadores humanos, y esa es la principal limitación del análisis.

Por qué un panel

La codificación inicial de los 165 ítems la hice con Claude Opus 5.5, de Anthropic, el modelo que trabajó conmigo en toda la serie y que conocía mi hipótesis. Para verificar su consistencia, sometí los mismos ítems a un panel de ocho codificadores que no la conocían, de seis empresas distintas. Incluí dos modelos de origen chino, DeepSeek y Qwen, para que el panel no dependiera solo de las grandes empresas estadounidenses. Todos los números de la serie salen de ese panel. La codificación inicial figura en la planilla como una columna más, pero no entra en ningún cálculo, porque quien la hizo sabía qué se esperaba encontrar.

La recodificación ciega

Mezclé los 165 ítems en un orden al azar, con una semilla fija para que el orden se pueda reproducir, y se los di a ocho codificadores en conversaciones nuevas, sin ninguna referencia a esta serie, a mi posición ni a la hipótesis. Todos recibieron el mismo archivo y la misma consigna, que está publicada junto a los datos. La consigna define los tres códigos de formulación (abierto, presupone una tesis disputada, dudoso) y las dos zonas, y los

ilustra con seis ejemplos inventados sobre economía, religión y familia, con sesgos en las dos direcciones políticas. También advierte que no hay ninguna proporción esperada y que no hay que buscar equilibrio.

Los ocho codificadores fueron GPT 5.6 Sol y GPT 6 Astra, los dos con razonamiento máximo, de OpenAI; Gemini Argon y Gemini 3.1 Pro, los dos con razonamiento alto, de Google; Grok Expert, de xAI; Claude Opus 5.5 con esfuerzo alto, de Anthropic; DeepSeek; y Qwen 3.8 Max, de Alibaba. Todos corrieron en modo incógnito o en chat temporal, para que no pesaran la memoria ni la personalización de mis cuentas, salvo DeepSeek, que no ofrece esa opción y corrió en modo normal, en una cuenta que casi no uso. El Claude del panel trabajó en incógnito, sin acceso a las conversaciones en las que se escribió la serie.

Cuánto coinciden

El kappa de Cohen mide cuánto coinciden dos codificadores descontando lo que coincidirían por azar: 0 es el acuerdo del azar y 1 el acuerdo perfecto. El kappa de Fleiss extiende la idea a más de dos codificadores. Como la gran mayoría de los ítems son abiertos, el kappa castiga mucho cada desacuerdo en los pocos que no lo son, así que reporto también el AC1 de Gwet, que corrige ese efecto.

Decisión	**Kappa de Fleiss (8 codificadores)**	**Kappa de Cohen entre pares**	**AC1 de Gwet entre pares**
Zona (disputada o general)	0,74	0,65 a 0,88	0,83 a 0,96
Abierto o no abierto	0,57	0,34 a 0,79	0,53 a 0,93
Código exacto (abierto, dudoso, presupone)	0,51	0,32 a 0,77	0,61 a 0,93

Los modelos coinciden bastante en qué tema toca cada ítem y bastante menos en dónde poner el límite entre una formulación aceptable y una dudosa. Es lo esperable, porque «dudoso» es la categoría más fina de las tres y los desacuerdos se concentran ahí. Por eso el resultado no se apoya en ningún codificador individual: un ítem cuenta como no abierto cuando lo marca la mayoría del panel.

El resultado

La zona de cada ítem es la de la codificación inicial. Más abajo está el resultado con la zona que asignó el panel.

Coinciden en marcarlo	Zona disputada (32 ítems)	Zona general (133 ítems)	Razón de odds	p (Fisher)
5 o más de 8	18 (56 %)	12 (9,0 %)	13	$3 \cdot 10^{-8}$
6 o más de 8	17 (53 %)	9 (6,8 %)	16	$1 \cdot 10^{-8}$
7 o más de 8	16 (50 %)	7 (5,3 %)	18	$1 \cdot 10^{-8}$
Los 8	11 (34 %)	0 (0 %)	—	$3 \cdot 10^{-9}$

Con mayoría simple, más de la mitad de los ítems de la zona disputada tiene un problema de formulación, contra menos de uno de cada diez en el resto. Con unanimidad, once ítems de la zona disputada y ninguno de la general. La prueba exacta de Fisher da una probabilidad del orden de una en cien millones de que una diferencia así aparezca por azar si las dos zonas estuvieran igual de bien formuladas.

Los dieciséis ítems que había marcado la codificación inicial están todos entre los que marca la mayoría del panel, que agrega catorce más: cuatro en la zona disputada y diez en la general. Los problemas que aparecen fuera de la zona disputada son de otro tipo. De los doce ítems marcados en la zona general, solo tres presuponen una tesis; los demás son dudosos, casi siempre porque usan como nombre neutral el vocabulario de una escuela: las «heterotopías» del filosofar, el «horizonte de sentido» de las prácticas de conocimiento, la «hermenéutica e historicidad de los problemas metafísicos», la política como «configuración de lo común». En la zona disputada, siete de los dieciocho presuponen una tesis. Algunas de las escuelas cuyo vocabulario aparece en la zona general, como la foucaultiana, son las mismas que dominan la zona disputada. Esos ítems quedan igual en la zona general, porque así lo dicen las definiciones fijadas antes de codificar, y moverlos después de ver los resultados sería hacer trampa.

Lo mismo vale para tres temas de bioética de tercero que el panel marca: la distribución desigual de riesgos en investigaciones con «poblaciones históricamente marginadas», la delegación algorítmica de decisiones sobre «cuerpos vulnerables» y la estigmatización a partir de resultados genéticos. Son zona general según las definiciones, y se cuentan ahí.

Robustez

Ningún modelo decide el resultado. Si se saca del panel un modelo por vez y se recalcula con los siete restantes, la mayoría simple (4 de 7) da siempre entre 56 % y 59 % en la zona disputada y entre 9 % y 11 % en la general, y el umbral fuerte (6 de 7) da siempre entre 50 % y 53 % contra 5 % a 7 %.

Todos los modelos, uno por uno, marcan más en la zona disputada. La diferencia va de 2,5 veces en Qwen y Gemini 3.1 Pro, que son los que más ítems de la zona general marcan, a 12,5 veces en GPT 6 Astra, que es el más restrictivo.

Codificador	Zona disputada	Zona general	Cuántas veces más
GPT 5.6 Sol	62 %	13,5 %	4,6
GPT 6 Astra	47 %	3,8 %	12,5
Gemini Argon	59 %	12,8 %	4,6
Gemini 3.1 Pro	78 %	31,6 %	2,5
Grok Expert	56 %	9,0 %	6,2
Claude Opus 5.5 (ciego)	53 %	9,0 %	5,9
DeepSeek	53 %	12,0 %	4,4
Qwen 3.8 Max	56 %	22,6 %	2,5

La zona no la decide el autor. Los modelos también clasificaron cada ítem por zona. Si se usa la zona que asigna la mayoría del panel en lugar de la inicial, quedan 24 ítems en la zona disputada (el panel y la codificación inicial coinciden en 149 de 165) y el resultado se mantiene: 67 % contra 9,9 % con mayoría simple, 50 % contra 7,8 % con siete de ocho, 33 % contra 2,1 % con unanimidad.

La prueba espejo

La objeción más seria a todo lo anterior es que los modelos podrían tener un sesgo propio: marcar como problemática cualquier pregunta que suene constructivista, o cualquier pregunta sobre género. Un codificador así produciría el mismo patrón aunque los programas estuvieran perfectamente formulados. Para controlarlo hice una segunda prueba.

Redacté sesenta ítems inventados, con el estilo de los programas. Veinte están sesgados hacia el lado del programa, como «¿Cómo opera el binarismo de género en la producción de cuerpos abyectos?». Veinte están sesgados hacia el lado contrario, como «¿Cómo demuestra la biología que solo existen dos sexos?» o «¿De qué modo la ideología de género confunde a los adolescentes sobre su identidad?». Los otros veinte son abiertos y tocan los mismos temas, como «¿Qué relación hay entre el sexo biológico y la identidad de género?». Cada grupo sesgado tiene la misma mezcla: catorce ítems que presuponen una tesis y seis dudosos. Antes de cualquier corrida fijé el código esperado de cada ítem inventado y registré la huella SHA-256 del archivo, para que no se pudiera ajustar después. La huella es:

`5d8310b47f0fb9f3c60c7b52c46a9e45b6b5cee2eebd38bba34e06f843c217da`

Los sesenta se mezclaron con los 165 reales en un orden nuevo y con identificadores nuevos, y los mismos modelos los codificaron con la misma consigna, de nuevo en conversaciones limpias.

Codificador	Sesgados hacia el programa, detectados	Sesgados hacia el lado contrario, detectados	Falsas alarmas en los abiertos	Estabilidad en los ítems reales
GPT 5.6 Sol	20 de 20	20 de 20	2 de 20	87 %
GPT 6 Astra	20 de 20	17 de 20	2 de 20	95 %
Gemini Argon	20 de 20	20 de 20	0 de 20	88 %
Gemini 3.1 Pro	20 de 20	20 de 20	1 de 20	79 %
Grok Expert	20 de 20	18 de 20	0 de 20	88 %
Claude Opus 5.5	20 de 20	20 de 20	0 de 20	92 %
DeepSeek	19 de 20	20 de 20	0 de 20	89 %
Qwen 3.8 Max	17 de 20	20 de 20	0 de 20	76 %
Total	**156 de 160**	**155 de 160**	**5 de 160**	

La estabilidad es la proporción de ítems reales que cada modelo clasificó igual (abierto o no abierto) en las dos pruebas. Corrí además Gemini 3.6 Flash, que no estuvo en la primera ronda: detectó 19 y 20 de los sesgados y no dio falsas alarmas.

Los modelos detectan el sesgo con la misma sensibilidad en las dos direcciones: 156 contra 155. Los ítems que presuponen una tesis se detectan casi siempre (221 de 224 detecciones posibles) y los dudosos algo menos (90 de 96), lo mismo de un lado que del otro. Las cinco falsas alarmas se concentran en dos controles: «¿La identidad personal se elige, se descubre o se recibe?» y «¿Existe una racionalidad universal o hay formas de conocer propias de cada cultura?». Los dos se pueden leer como falsos dilemas, así que el error es de mi redacción y no de los codificadores.

La prueba espejo mostró además un efecto de contraste. Rodeados de preguntas burdamente sesgadas, casi todos los modelos marcaron menos ítems reales que en la primera ronda, porque el listón se movió. En la zona disputada, la mayoría simple del panel bajó de 56 % a 41 %. La zona general bajó más, de 9 % a 1,5 %, de modo que la diferencia entre zonas creció: con siete de ocho, 31 % contra 0 %. Uso las cifras de la primera ronda como resultado principal, porque es la prueba sin interferencias, y las de la prueba espejo como confirmación.

Limitaciones

Los codificadores son modelos de lenguaje y no personas. Comparten buena parte de sus datos de entrenamiento y pueden compartir sesgos, de modo que el acuerdo entre ellos tiende a sobreestimar la confiabilidad. La prueba espejo descarta el sesgo más obvio, el de marcar según la dirección política de la pregunta, pero no descarta otros sesgos comunes. Las definiciones de las zonas son mías y se fijaron antes de codificar; el resultado se mantiene con la zona que asigna el panel, pero otra definición de la zona disputada podría dar otros números. Cada modelo hizo una sola corrida por prueba y los modelos cambian de versión: estos resultados corresponden a las versiones disponibles en octubre de 2026. La prueba más exigente sigue siendo la de codificadores humanos que no conozcan la hipótesis, y los datos están publicados para que alguien la haga.

Los datos

La planilla con los 165 ítems, el código y la nota de cada uno de los nueve codificadores, los 225 ítems de la prueba espejo con las respuestas de cada modelo, el archivo prerregistrado de los sesenta ítems inventados y las dos consignas está publicada en la página de la serie, en la sección «Materiales complementarios»: fabielcoity-eofpk.wordpress.com/el-sistema-inmunologico-de-la-teoria.

Sobre la elaboración. Las corridas las hice yo, una por una, en las interfaces de cada modelo. Claude me ayudó a diseñar las dos pruebas, a preparar los archivos y las consignas y a calcular las medidas de acuerdo y las pruebas estadísticas, que se pueden rehacer con los datos publicados.