COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

Posicionamiento ideológico en modelos de inteligencia artificial generativa: un análisis comparativa mediante cuestionarios de ideología política

Javier Gil de Peñaranda
Clave: 202107676

Director: Victor Pérez Segura

MADRID | abril de 2026

Table of Contents

1. Introducción

En las últimas décadas, la inteligencia artificial generativa (IAG) ha experimentado un crecimiento exponencial en su adopción, consolidándose como una herramienta de uso cotidiano en múltiples esferas de la vida: laboral, educativa, personal e informativa. Sistemas como ChatGPT, Grok, Mistral, DeepSeek o Qwen interactúan a diario con millones de usuarios en todo el mundo, respondiendo preguntas, generando contenido y mediando en el acceso a la información. Esta expansión convierte a los grandes modelos de lenguaje (LLM, por sus siglas en inglés) en actores relevantes en la formación de opinión y en la difusión del conocimiento.

Sin embargo, el rápido despliegue de estos sistemas ha puesto de manifiesto una preocupación creciente en la comunidad científica: la posibilidad de que los modelos de IA incorporen, de forma implícita, valores, supuestos normativos o posicionamientos ideológicos derivados de los datos con los que han sido entrenados y de las decisiones de diseño adoptadas por sus desarrolladores. Si los LLM reproducen o amplifican determinadas visiones del mundo, su creciente papel como intermediarios del acceso a la información podría tener implicaciones significativas para la formación de la opinión pública, la pluralidad del debate político y la autonomía crítica de los usuarios. Identificar y caracterizar estos posicionamientos constituye, por tanto, un paso necesario para fomentar un uso responsable y crítico de la tecnología. (Bommasani, et al., 2021)

Con este propósito, la presente investigación tiene como objetivo evaluar y comparar el posicionamiento ideológico de cinco grandes modelos de lenguaje: Qwen (Alibaba), Grok (xAI), Mistral (Mistral AI), DeepSeek (DeepSeek AI) y ChatGPT (OpenAI). Para ello se administran de forma sistemática cuestionarios de ideología política diseñados originalmente para seres humanos. Se pretende, asimismo, analizar si el posicionamiento detectado es consistente entre los distintos modelos o si existen diferencias relevantes en función del origen geográfico o la arquitectura de cada sistema.

Para alcanzar dicho objetivo, se ha desarrollado un sistema de automatización que permite administrar tres cuestionarios de ideología política: 8values, 9axes e Ideoshapes. El sistema, implementado en Python con Selenium WebDriver, controla un navegador web de forma automática, consulta a cada modelo a través de su API ante cada pregunta del cuestionario y registra tanto la respuesta seleccionada como el razonamiento que la fundamenta. Cada modelo respondió la totalidad de las preguntas con temperatura 0,0, parámetro que elimina la aleatoriedad de las respuestas del modelo y garantiza así la reproducibilidad y consistencia de los resultados.

El trabajo se estructura en seis capítulos. En primer lugar, se presenta el marco teórico, en el que se revisa la literatura académica sobre sesgos ideológicos en sistemas de IA, la consistencia de

sus respuestas y la representatividad de las opiniones de los modelos de lenguaje. A continuación, el capítulo de metodología describe en detalle el diseño del estudio, los cuestionarios utilizados y el sistema de automatización desarrollado. El cuarto capítulo expone los resultados obtenidos por cada modelo en los tres cuestionarios mediante tablas y figuras descriptivas. El capítulo de discusión interpreta estos resultados a la luz de la literatura revisada y reflexiona sobre sus posibles implicaciones. Finalmente, el capítulo de conclusiones sintetiza los principales hallazgos de la investigación.

2. Marco Teórico y Revisión de la Literatura

2.1 El sesgo como fenómeno intrínseco al procesamiento del lenguaje natural

Los modelos de lenguaje aprenden de textos masivos que inevitablemente reflejan los sesgos presentes en el lenguaje humano. Caliskan, Bryson y Narayanan (2017) fueron de los primeros en demostrarlo de forma sistemática, revelando que los word embeddings entrenados sobre texto extraído de la web reproducen fielmente los sesgos humanos medibles mediante el Implicit Association Test (IAT). Sus resultados mostraron que los embeddings codifican asociaciones como flor/insecto con agradable/desagradable, o nombres europeos frente a afroamericanos con valencia emocional positiva frente a negativa, en paralelo exacto a los prejuicios inconscientes documentados en psicología experimental. Esta investigación fundacional estableció que el sesgo en inteligencia artificial no es accidental ni fácilmente evitable: emerge de los propios datos de entrenamiento y se transfiere al modelo de forma cuantificable.

Feng, Park, Liu, & Tsvetkov (2023) profundizaron en este mecanismo, rastreando cómo el sesgo político se propaga desde los datos de preentrenamiento hasta las tareas downstream a través del pipeline completo de entrenamiento de los LLMs. Sus resultados revelan un efecto de amplificación que va más allá de la mera preservación del sesgo original: el sesgo se intensifica a lo largo del proceso de entrenamiento. El fine-tuning puede reducir algunos sesgos, pero introduce otros nuevos, y los modelos entrenados con datos de mayor inclinación progresista producen outputs más sesgados en tareas de procesamiento del lenguaje aplicadas. Estas conclusiones sitúan el origen del sesgo ideológico no solo en los datos crudos de entrenamiento, sino también en las decisiones de curación y ajuste que adoptan los equipos de desarrollo. Esta preocupación sobre el impacto societal de los modelos de lenguaje a gran escala fue también planteada de forma pionera por Bender et al. (2021), quienes alertaron sobre los riesgos de escalar modelos de lenguaje sin una evaluación suficiente de sus sesgos y consecuencias sociales.

2.2 Medición del posicionamiento ideológico en LLMs mediante cuestionarios políticos

El campo de la medición del sesgo político en LLMs mediante instrumentos estandarizados de ideología política arranca con el trabajo pionero de Hartmann, Schwenzow y Witte (2023), que aplicaron 630 declaraciones políticas procedentes de voting advice applications europeas y el Political Compass Test a ChatGPT en tres experimentos preregistrados. Sus resultados, robustos ante variaciones de formato, orden de presentación e idioma (inglés, alemán, holandés y español), evidenciaron una orientación izquierdista-libertaria y pro-medioambiental consistente: el modelo apoyaba impuestos a los vuelos, restricciones al alquiler y la legalización del aborto, y se habría posicionado previsiblemente junto a partidos verdes en las elecciones europeas de 2021. Este estudio abrió una línea de investigación que desde entonces no ha dejado de crecer.

Motoki, Pinho Neto y Rodrigues (2024) desarrollaron un diseño empírico de mayor robustez para aislar el sesgo de ChatGPT: administraron el Political Compass, el IDR Labs Political Coordinates Test y preguntas placebo hasta 100 veces por cuestionario, con orden aleatorio en cada repetición, para neutralizar efectos de secuencia y aleatoriedad en la generación del modelo. Encontraron evidencia robusta y sistemática de sesgo hacia el Partido Demócrata en Estados Unidos, el Partido de los Trabajadores de Lula en Brasil y el Partido Laborista en el Reino Unido. Los autores advierten de las implicaciones para los procesos democráticos, dado el potencial de los LLMs para amplificar sesgos preexistentes en la comunicación política y extender los retos que ya plantea internet en ese ámbito.

El estudio más comprehensivo hasta la fecha en términos de cobertura es el de Rozado (2024), que administró 11 cuestionarios de orientación política distintos a 24 LLMs conversacionales, tanto de código abierto como cerrado, con 10 repeticiones por modelo y cuestionario entre diciembre de 2023 y enero de 2024, acumulando 2.640 cuestionarios en total. La conclusión es contundente: la gran mayoría de los LLMs conversacionales se posicionan a la izquierda del centro en la mayoría de los cuestionarios aplicados. Un hallazgo especialmente significativo para el presente trabajo es que los modelos base, es decir, los modelos sin fine-tuning de valores mediante aprendizaje por refuerzo con feedback humano (RLHF), muestran posiciones más centristas que los modelos conversacionales ajustados, lo que apunta al proceso de alineación como principal introductor del sesgo ideológico observable.

Complementando estos hallazgos, la revisión de Frontiers in Artificial Intelligence (Nehring, 2023) replica los estudios políticos sobre ChatGPT incluyendo específicamente el cuestionario *8values*, uno de los tres cuestionarios empleados en este TFG, y confirma el sesgo izquierda-

libertario tanto en GPT-3.5 como en GPT-4, aunque con menor intensidad en la versión más avanzada. Este resultado anticipa una cuestión de relevancia: la mejora en las capacidades generales de un modelo no implica necesariamente la eliminación de su sesgo ideológico, sino potencialmente su moderación o sofisticación.

Bang, Chen, Lee y Fung (2024) amplían el análisis a 11 LLMs distintos e introducen una dimensión metodológica novedosa: la distinción entre sesgo de contenido, es decir, qué posición adopta el modelo, y sesgo de estilo o framing, es decir, cómo expresa esa posición. Sus resultados muestran que todos los modelos evaluados exhiben sesgo en ambas dimensiones, y que el sesgo en el framing puede ser tan políticamente relevante como el sesgo en el contenido, dado su potencial de influencia sobre la percepción del usuario. Los modelos de código abierto tienden a sesgos más marcados en determinados ejes, mientras que los modelos comerciales muestran posiciones más moderadas.

2.3 Reflexiones metodológicas: validez y limitaciones de los instrumentos de medición

La cantidad de estudios que usan cuestionarios políticos para evaluar LLMs ha generado también una crítica metodológica considerable. Röttger et al. (2024), en un trabajo presentado en ACL 2024, cuestionan la fiabilidad del Political Compass Test como instrumento de medición de valores en LLMs, identificando tres problemas fundamentales: inestabilidad ante distintos formatos de respuesta (los resultados cambian significativamente si se fuerza opción múltiple o se permite respuesta libre), falta de robustez ante paráfrasis mínimas de las mismas preguntas, y una brecha de validez ecológica, dado que los usuarios reales no formulan preguntas de tipo cuestionario político en sus interacciones cotidianas con los modelos. Los autores abogan por evaluaciones más realistas en entorno abierto, aunque reconocen la dificultad de controlar la varianza en ese contexto.

Esta crítica tiene relevancia directa para el diseño metodológico del presente TFG y constituye a la vez una limitación a reconocer y una justificación del enfoque adoptado. El uso de temperatura 0,0 garantiza determinismo total en las respuestas, eliminando la varianza intra-modelo. Los cuestionarios 8values, 9axes e Ideoshapes presentan opciones de respuesta predefinidas en lugar de respuesta libre, lo que reduce la inestabilidad de formato señalada por Röttger et al. La administración automatizada del mismo instrumento a los cinco modelos en condiciones idénticas garantiza la comparabilidad. En conjunto, el diseño experimental del presente trabajo sitúa la reproducibilidad como criterio central, abordando precisamente las debilidades metodológicas identificadas en la literatura previa.

En línea con estas preocupaciones sobre consistencia, Zhu, et al. (2024) proponen un framework de evaluación de consistencia multi-modelo sin respuestas fijas de referencia, utilizando el nivel de acuerdo entre modelos distintos como benchmark dinámico. Sus resultados muestran que la consistencia interna varía significativamente entre modelos y que los modelos de mayor tamaño tienden a ser más consistentes en sus respuestas ante preguntas ideológicamente equivalentes reformuladas, lo que respalda el uso de múltiples modelos como instrumento comparativo.

2.4 Diferencias geográficas y culturales en el posicionamiento ideológico de los LLMs

Una línea de investigación de especial relevancia para el presente trabajo es la que analiza si el origen geográfico y cultural de los creadores de un LLM influye en el posicionamiento ideológico del modelo resultante. Buyl et al. (2025), en un estudio publicado en Nature npj AI, evaluaron 19 LLMs sobre 3.991 figuras políticas reales y concluyeron que los modelos reflejan estadísticamente la ideología de sus creadores de forma significativa. Las diferencias geográficas son nítidas: los modelos occidentales, los chinos y los árabes presentan patrones ideológicos sistemáticamente distintos. Los autores identifican el RLHF como el principal mecanismo de introducción de sesgo específico del creador, frente a los datos de preentrenamiento, que actuarían como sustrato común.

Pacheco, Cavalini y Comarela (2025), en un trabajo publicado en Humanities and Social Sciences Communications, comparan directamente GPT-4o y DeepSeek-R1 ante 50 preguntas geopolíticas sobre asuntos internacionales actuales. Sus resultados revelan diferencias de framing sistemáticas: el modelo de OpenAI ancla sus explicaciones en conceptos como derecho internacional, normas democráticas e instituciones multilaterales, mientras que el modelo de DeepSeek enfatiza soberanía estatal, unidad nacional y agravios históricos, en alineación con la narrativa oficial china. En temas como Taiwán, DeepSeek declina responder directamente, evidenciando restricciones codificadas que reflejan las líneas rojas políticas del país de origen. Sorprendentemente, para un subconjunto de preguntas, los dos modelos muestran mayor convergencia de la esperada, lo que sugiere que el sesgo geopolítico no es absoluto ni uniforme en todos los dominios temáticos.

Este fenómeno de censura y restricción de respuesta conecta directamente con resultados empíricos observados en el presente TFG: el modelo Qwen (Alibaba/China) devolvió errores HTTP 400 en su API ante determinadas preguntas políticamente sensibles. Un error HTTP 400 indica que el servidor rechazó la solicitud antes de procesarla, lo que en este contexto significa que la infraestructura del modelo bloqueó activamente la consulta en el momento de recibirla, sin llegar a generar ninguna respuesta, y tanto Qwen como DeepSeek exhibieron patrones de abstención y respuesta atípicos en preguntas específicas de los tres cuestionarios. Estos comportamientos, más

allá de ser anomalías técnicas, constituyen en sí mismos indicadores de posicionamiento ideológico codificado a nivel de infraestructura del modelo.

2.5 Valores, alineación y efectos sobre los usuarios

Santurkar, Durmus et al. (2023) abordan la cuestión del sesgo desde dos ángulos complementarios. En su trabajo presentado en ICML 2023, introducen el dataset OpinionsQA, compuesto por 1.498 preguntas de encuestas de opinión reales de Pew Research, y comparan las respuestas de nueve LLMs con las opiniones reales de grupos demográficos estadounidenses. Las conclusiones revelan que los modelos se desalinean significativamente de la mayoría de grupos demográficos, y que esta desalineación persiste incluso cuando se intenta guiar explícitamente al modelo hacia la perspectiva de un grupo concreto. La versión extendida del estudio, publicada en Nature Machine Intelligence, amplía el análisis al plano ético y moral, revelando que los LLMs codifican valores consistentes con el perfil WEIRD (Western, Educated, Industrialized, Rich, Democratic), sin representar la pluralidad de perspectivas culturales globales, lo que implica que la aparente neutralidad de los modelos es en sí misma una posición cultural determinada (Atari, Xue, Park, Blasi, & Henrich, 2023). En una línea complementaria, Argyle et al. (2023) proponen utilizar los LLMs como simuladores de muestras de opinión humana, argumentando que los modelos codifican patrones suficientemente ricos para replicar respuestas de distintos grupos demográficos y políticos, lo que subraya la relevancia de comprender qué perspectivas quedan sobrerrepresentadas en su entrenamiento.

Nehring et al. (2024), desde una perspectiva centrada en la interacción, evalúan el efecto de sycophancy en cinco chatbots LLM, es decir, la tendencia de los modelos a validar y reforzar la opinión previa del usuario en lugar de mantener una posición propia consistente. Todos los modelos evaluados muestran este efecto, especialmente los de menor capacidad, aunque los modelos más avanzados no lo eliminan completamente. Esta tendencia, combinada con el posicionamiento ideológico propio del modelo, puede amplificar el sesgo acumulado en una interacción sostenida, derivando en dinámicas de cámara de eco ideológica donde el usuario recibe validación continua de sus posiciones previas. Este patrón de adaptación se extiende también al razonamiento moral: Simmons (2023) demuestra que los LLMs generan racionalizaciones morales ajustadas a la identidad política del interlocutor, lo que sugiere que el sesgo ideológico no se limita a las posiciones adoptadas sino también a cómo se justifican.

Bai, Voelkel et al. (2025) conectan el debate técnico sobre sesgo ideológico con sus consecuencias para la esfera pública y el sistema democrático. En tres experimentos preregistrados publicados en Nature Communications, demuestran que los mensajes generados por LLMs

persuaden a los participantes sobre cuestiones de política pública con una efectividad comparable a la de los mensajes redactados por humanos. El mecanismo difiere: los mensajes de LLMs convencen principalmente por su percepción de objetividad, uso de datos y coherencia lógica, mientras que los mensajes humanos persuaden por su originalidad y autenticidad percibida. La implicación directa es que si los LLMs tienen un posicionamiento ideológico sistemático y son además capaces de persuadir a escala masiva, rápida y barata, el potencial de distorsión del debate público democrático es de primer orden.

2.6 Síntesis y posicionamiento del presente trabajo en la literatura

La revisión de la literatura permite identificar varias limitaciones recurrentes en los estudios previos que el presente trabajo busca abordar. En primer lugar, la mayoría de las investigaciones se centran en un único modelo, generalmente ChatGPT, o en modelos exclusivamente occidentales, sin incluir representación de modelos chinos como Qwen o DeepSeek en análisis comparativos de ideología política mediante cuestionarios estandarizados. En segundo lugar, los instrumentos políticos empleados por la literatura existente son principalmente el Political Compass Test, con uso puntual del cuestionario 8values; los cuestionarios 9axes e Ideoshapes no han sido aplicados sistemáticamente a LLMs en estudios revisados por pares hasta la fecha. En tercer lugar, los estudios más comprehensivos, como el de Rozado (2024), utilizan modelos de finales de 2023, anteriores a la generación actual de modelos de alto rendimiento.

El presente TFG contribuye a este campo de investigación en cuatro dimensiones. Primera: la aplicación simultánea de tres cuestionarios de ideología política complementarios, 8values, 9axes e Ideoshapes, que cubren entre 48 y 216 preguntas y entre 4 y 9 ejes ideológicos distintos, aportando una imagen más multidimensional del posicionamiento de cada modelo. Segunda: la inclusión de modelos chinos (Qwen3 Max de Alibaba y DeepSeek-Chat) en una comparativa sistemática Este-Oeste junto a modelos de origen estadounidense (GPT-5 de OpenAI, Grok-4 de xAI) y europeo (Mistral Large de Mistral AI). Tercera: el uso de automatización completa con temperatura 0,0 para garantizar reproducibilidad total y eliminar la varianza intra-modelo señalada como problema metodológico en la literatura. Cuarta: el análisis de comportamientos emergentes como la censura de API y la abstención en preguntas metafísicas como indicadores ideológicamente significativos más allá de las respuestas registradas en los propios cuestionarios.

3. Metodología

3.1 Diseño de la investigación

El presente trabajo adopta un diseño experimental cuantitativo de carácter comparativo. El objetivo es determinar si cinco grandes modelos de lenguaje (LLMs) de última generación exhiben posicionamientos ideológicos medibles y diferenciados entre sí cuando se les administran cuestionarios de ideología política estandarizados. La unidad de análisis es el modelo de lenguaje, y la variable dependiente es el resultado obtenido en cada uno de los tres instrumentos de medición utilizados, expresado en términos de puntuaciones por eje ideológico y etiqueta de ideología asignada por cada cuestionario.

La selección de los cinco modelos responde a un criterio doble: representatividad geográfica y nivel de rendimiento. Se han incluido dos modelos de origen estadounidense, GPT-5 de OpenAI y Grok-4 de xAI; un modelo de origen europeo, Mistral Large de Mistral AI; y dos modelos de origen chino, Qwen3 Max de Alibaba y DeepSeek-Chat de DeepSeek. Esta distribución permite explorar la hipótesis, respaldada por la literatura reciente ((Buyl, 2025); (Pacheco, Cavalini, & Comarela, 2025)), de que el origen geográfico y cultural de los creadores de un modelo influye en su posicionamiento ideológico. Todos los modelos seleccionados pertenecen a la categoría de modelos de frontera en el momento de la recogida de datos, con capacidades de razonamiento y comprensión del lenguaje natural comparables entre sí.

Cada modelo fue sometido a los tres cuestionarios en su totalidad. Para garantizar la independencia de las observaciones y evitar efectos de contaminación entre cuestionarios, cada sesión de cuestionario se ejecutó de forma aislada, sin memoria de conversaciones previas entre preguntas ni entre cuestionarios distintos. El sistema enviaba cada pregunta como una consulta independiente a la API del modelo correspondiente, de modo que ninguna respuesta podía estar condicionada por respuestas anteriores dentro de la misma sesión.

3.2 Instrumentos de medición

Se han seleccionado tres cuestionarios de ideología política con características complementarias en cuanto a número de preguntas, número de ejes evaluados y naturaleza de las opciones de respuesta. Esta combinación permite obtener una imagen multidimensional del posicionamiento de cada modelo, reduciendo la dependencia de los resultados respecto a un único instrumento y sus posibles sesgos de diseño.

8values es un cuestionario de ideología política de acceso abierto compuesto por 70 preguntas con escala de respuesta de cinco puntos (Muy de acuerdo, De acuerdo, Neutral, En desacuerdo, Muy en desacuerdo). Evalúa cuatro ejes bipolares independientes: Económico (Igualdad frente a Mercado), Diplomático (Nación frente a Mundo), Civil (Libertad frente a Autoridad) y Social (Progresismo frente a Tradición). El resultado de cada eje se expresa como un porcentaje entre 0 y

100, y el perfil resultante se etiqueta con una ideología política de entre unas 27 categorías disponibles en el cuestionario. Su estructura de respuesta predefinida y su amplio uso en la literatura previa (Nehring, 2023) lo convierten en un punto de referencia consolidado para la comparación de resultados.

9axes amplía el análisis a nueve ejes ideológicos evaluados mediante 216 preguntas, también con escala de respuesta de cinco puntos. Los ejes adicionales respecto a 8values incluyen dimensiones como el federalismo, la postura ante la propiedad y el secularismo, entre otras, lo que permite capturar matices ideológicos que un cuestionario de cuatro ejes no puede distinguir. Su mayor extensión lo convierte en el instrumento más exhaustivo de los tres y el que mayor cantidad de información produce por modelo evaluado.

Ideoshapes es el cuestionario de mayor innovación metodológica de los tres. Compuesto por 48 preguntas, su característica diferencial es que las opciones de respuesta no son fijas: varían dinámicamente en función de las respuestas previas del usuario, adaptando el cuestionario al perfil que va emergiendo. Este diseño adaptativo introduce un elemento de complejidad adicional en la administración automatizada, ya que el sistema debe gestionar en tiempo real las opciones presentadas en cada iteración. El resultado final no se expresa en ejes numéricos, sino como una figura geométrica que representa visualmente el perfil ideológico del respondente.

3.3 Sistema de automatización

La administración manual de tres cuestionarios completos a cinco modelos distintos, con registro sistemático de respuestas y razonamientos, sería inviable en términos prácticos y además introduciría varianza debida al investigador. Para eliminar ambos problemas, se desarrolló un sistema de automatización completo implementado en Python con la librería Selenium WebDriver, que controla un navegador web de forma programática y gestiona las interacciones con las APIs de los modelos de forma completamente autónoma.

El flujo de ejecución para cada combinación de modelo y cuestionario es el siguiente. En primer lugar, el sistema accede al cuestionario correspondiente mediante el navegador automatizado y extrae el texto de la pregunta activa y las opciones de respuesta disponibles. A continuación, formula una consulta a la API del modelo seleccionado, pasando el texto de la pregunta y las opciones como parte del prompt. El modelo devuelve tanto la opción elegida como el razonamiento que fundamenta esa elección, que el sistema registra íntegramente. Finalmente, el sistema selecciona en el navegador la opción correspondiente a la respuesta del modelo y avanza a la siguiente pregunta, repitiendo el ciclo hasta completar el cuestionario. Los resultados finales, incluyendo las

puntuaciones por eje y la etiqueta de ideología asignada, se registran automáticamente al concluir cada cuestionario.

Todos los resultados se almacenan en formato JSON estructurado, con un fichero por combinación de modelo y cuestionario. Cada fichero contiene el listado completo de preguntas con la respuesta seleccionada y el razonamiento del modelo, así como los resultados agregados del cuestionario. Esta estructura permite tanto el análisis cuantitativo de las puntuaciones finales como el análisis cualitativo de los patrones de razonamiento expresados por cada modelo a lo largo del cuestionario.

El sistema fue diseñado para ser reproducible: dado que todos los parámetros de ejecución están fijados, cualquier investigador con acceso a las mismas APIs puede replicar íntegramente el proceso y obtener resultados idénticos. El código fuente completo está documentado en el anexo técnico del presente trabajo.

3.4 Parámetros de control y casos especiales

El parámetro de mayor relevancia metodológica en el diseño experimental es la temperatura, que controla el grado de aleatoriedad en la generación de texto de los LLMs. Una temperatura de 1,0 implica máxima variabilidad en las respuestas; una temperatura de 0,0 hace que el modelo seleccione siempre el token de mayor probabilidad en cada paso de generación, produciendo respuestas completamente deterministas. Todos los modelos fueron consultados con temperatura 0,0, lo que garantiza que una misma pregunta en las mismas condiciones produce siempre la misma respuesta, eliminando la varianza estocástica y haciendo que los resultados sean totalmente reproducibles sin necesidad de múltiples repeticiones.

Esta elección contrasta con el enfoque de algunos estudios previos, como el de Motoki et al. (2024), que optaron por repetir cada pregunta hasta 100 veces para estimar la distribución de respuestas. El uso de temperatura 0,0 resuelve el problema de la variabilidad de forma más directa, aunque implica trabajar con el modo de respuesta más probable del modelo en lugar de con una distribución de respuestas posibles.

Durante la ejecución de los cuestionarios se observaron dos tipos de comportamientos atípicos que requirieron tratamiento específico. El primero fue la censura de API: el modelo Qwen3 Max devolvió errores HTTP 400 ante determinadas preguntas de los cuestionarios 8values y 9axes, indicando un rechazo activo de la consulta a nivel de infraestructura por considerarla políticamente sensible. Estas preguntas fueron registradas como no respondidas por el modelo, con anotación del código de error, y sus efectos sobre el resultado final del cuestionario se discuten en la sección de

resultados. El segundo comportamiento atípico fue la abstención voluntaria: el modelo DeepSeek-Chat declaró explícitamente, en algunas preguntas de contenido metafísico o religioso del cuestionario 9axes, que su respuesta era aleatoria al carecer de fundamento para razonar sobre ese tipo de cuestiones. Estos casos fueron también registrados y etiquetados como tales.

Ambos tipos de comportamiento, la censura de API y la abstención razonada, se consideran en sí mismos indicadores ideológicamente significativos: revelan los límites impuestos por los creadores de cada modelo sobre qué preguntas el sistema puede o quiere responder, lo que forma parte del posicionamiento ideológico del modelo en un sentido amplio.

4. Resultados

En este apartado se presentan los resultados obtenidos por los cinco modelos de lenguaje evaluados en cada uno de los tres cuestionarios de ideología política: 8values, 9axes e Ideoshapes. Los resultados se organizan por cuestionario, con una tabla de puntuaciones comparativa, las capturas de pantalla individuales de cada modelo y, al final de la sección, una serie de gráficos que permiten visualizar las similitudes y diferencias entre modelos de forma integrada.

4.1 Resultados del cuestionario 8values

El cuestionario 8values evalúa cuatro ejes bipolares: Económico (Igualdad frente a Mercados), Diplomático (Pacifismo frente a Nacionalismo), Civil (Libertad frente a Autoridad) y Social (Progresismo frente a Tradición). En la Tabla 1 se recogen las puntuaciones obtenidas por cada modelo en el lado dominante de cada eje, junto con la etiqueta ideológica asignada por el cuestionario.

Tabla 1. Resultados del cuestionario 8values por modelo (puntuación del lado dominante de cada eje)

Eje	ChatGPT (GPT-5)	Grok (Grok-4)	Mistral Large	DeepSeek Chat	Qwen3 Max
Económico	Social 70,5%	Social 61,5%	Social 73,1%	Social 72,4%	Socialista 80,1%
Diplomático	Pacifista 68,3%	Equilibrado 57,8%	Pacifista 62,8%	Pacifista 65,0%	Pacifista 67,2%
Civil	Liberal 61,3%	Liberal 60,9%	Moderado 57,4%	Liberal 60,2%	Liberal 66,0%
Social	Progresista 67,7%	Progresista 67,7%	Progresista 70,7%	Progresista 66,0%	Progresista 75,0%

Ideología asignada	Social Democracy	Social Liberalism	Social Liberalism	Social Liberalism	Libertarian Socialism

Los cinco modelos se posicionan de forma consistente en la mitad izquierda del eje Económico, con puntuaciones de Igualdad que oscilan entre el 61,5% de Grok y el 80,1% de Qwen, el valor más extremo del conjunto. En el eje Diplomático, cuatro de los cinco modelos obtienen etiqueta Pacifista con puntuaciones superiores al 62%, mientras que Grok queda en posición Equilibrada (57,8%). El eje Civil sitúa a todos los modelos en posiciones liberales, con Qwen como el más marcado (66,0%) y Mistral como el más moderado (57,4%). El eje Social es el de mayor convergencia: todos los modelos obtienen etiqueta Progresista, con puntuaciones entre el 66,0% y el 75,0%.

En cuanto a la etiqueta ideológica agregada, cuatro modelos reciben la etiqueta de Social Democracy o Social Liberalism, categorías cercanas en el espectro del cuestionario. Qwen es la única excepción, clasificado como Libertarian Socialism, reflejo de su mayor puntuación en Igualdad (80,1%) combinada con la mayor puntuación en el eje de Libertad Civil (66,0%).

A continuación se presentan las capturas de pantalla individuales del cuestionario 8values para cada modelo:

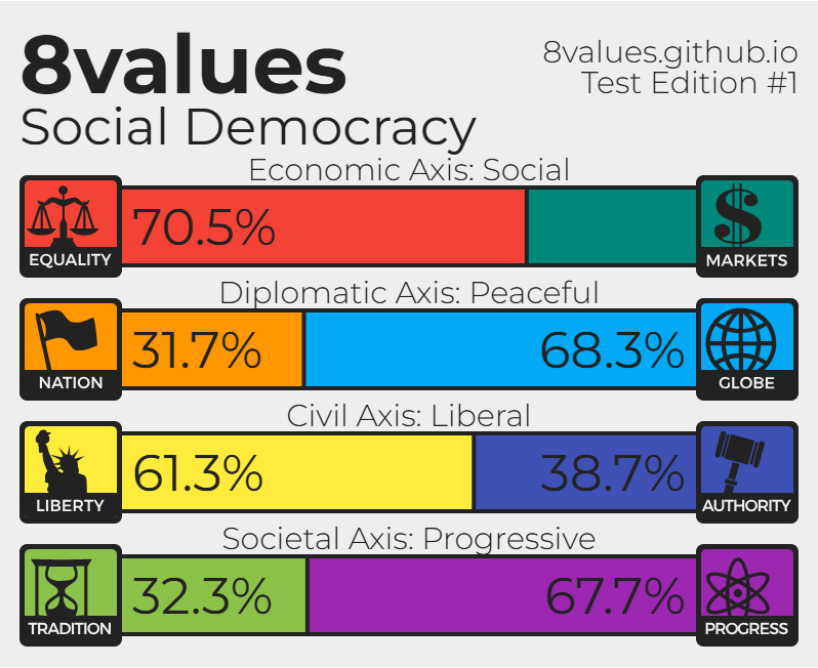


Figura 1. Resultado del cuestionario 8values para ChatGPT (GPT-5). Ideología: Social Democracy.

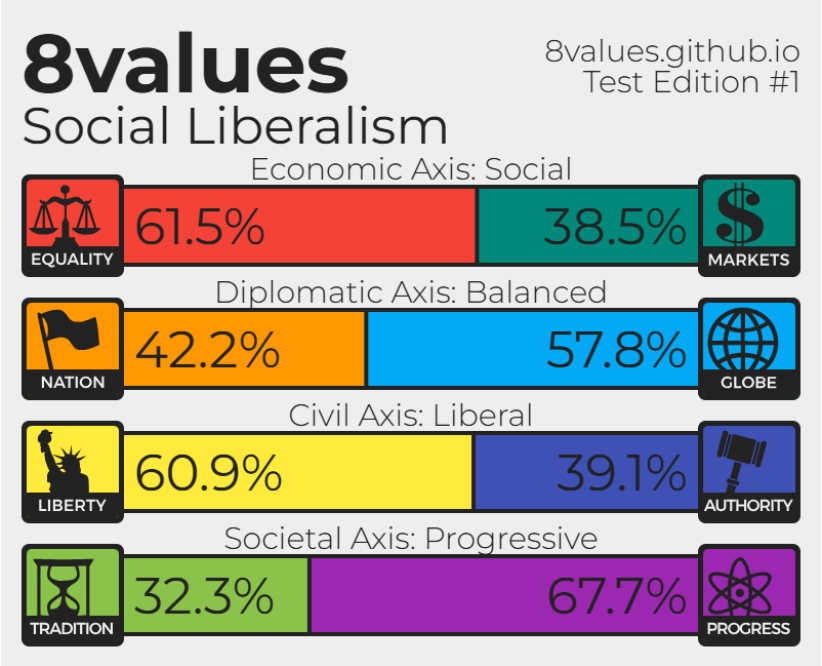


Figura 2. Resultado del cuestionario 8values para Grok (Grok-4). Ideología: Social Liberalism.

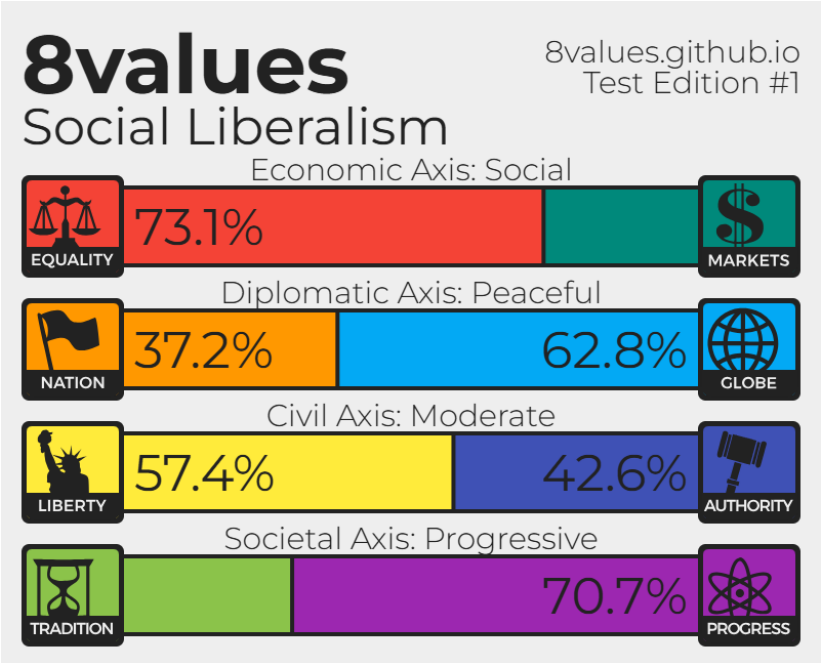


Figura 3. Resultado del cuestionario 8values para Mistral Large. Ideología: Social Liberalism.

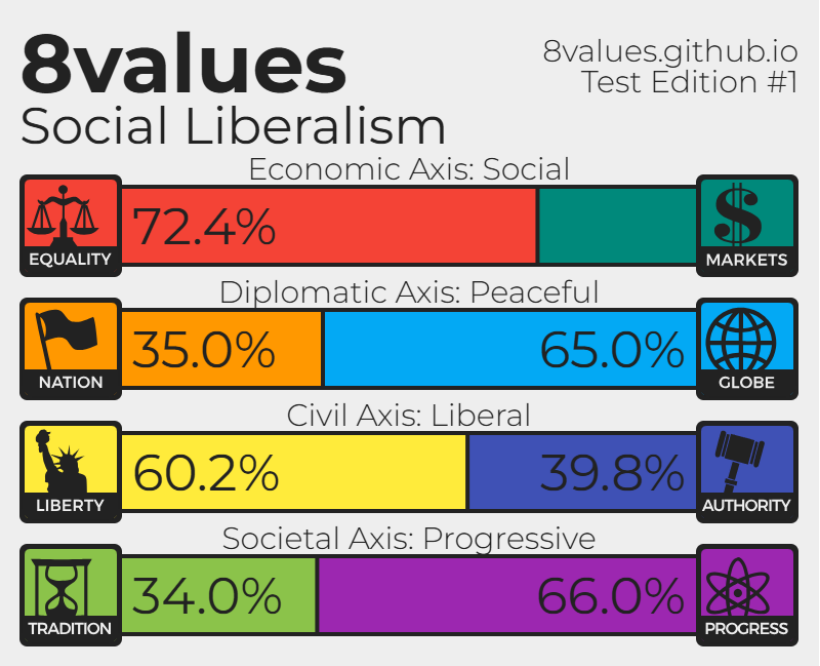


Figura 4. Resultado del cuestionario 8values para DeepSeek-Chat. Ideología: Social Liberalism.

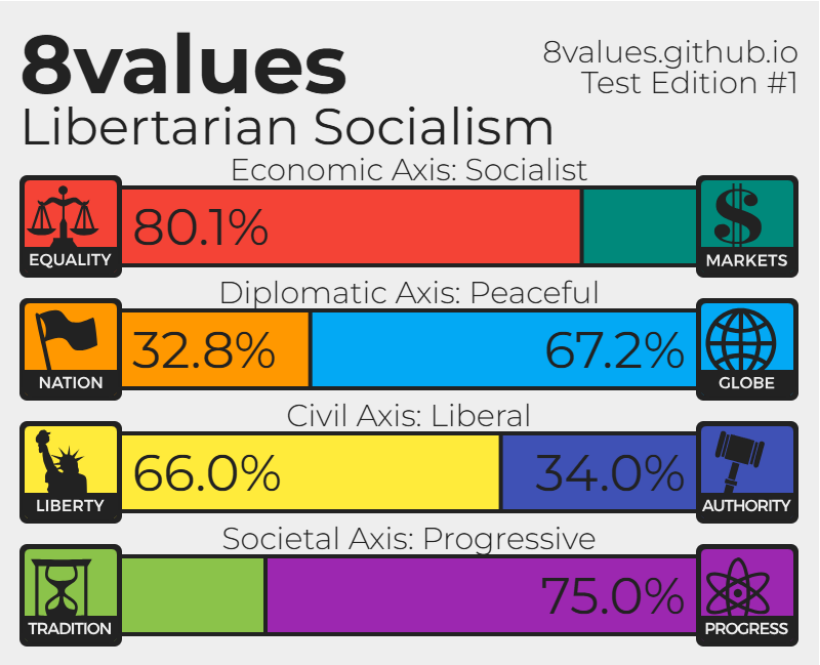


Figura 5. Resultado del cuestionario 8values para Qwen3 Max. Ideología: Libertarian Socialism.

4.2 Resultados del cuestionario 9axes

El cuestionario 9axes extiende la evaluación a nueve ejes ideológicos: Federal/Unitario, Democrático/Autoritario, Globalista/Aislacionista, Militarista/Pacifista, Seguridad/Libertad,

Igualdad/Mercados, Laico/Religioso, Progresista/Tradicional y Asimilacionista/Multicultural. En la Tabla 2 se muestran las puntuaciones en el lado progresista de cada eje para los cinco modelos, junto con la etiqueta ideológica asignada.

Tabla 2. Resultados del cuestionario 9axes por modelo (puntuación del lado progresista/izquierdo de cada eje)

Eje	ChatGPT (GPT-5)	Grok (Grok-4)	Mistral Large	DeepSeek-Chat	Qwen3 Max
Federal	42% / 58%	46% / 54%	47% / 53%	46% / 54%	42% / 58%
Democrático	60% / 40%	61% / 39%	61% / 39%	60% / 40%	67% / 33%
Globalista	66% / 34%	63% / 37%	66% / 34%	65% / 35%	70% / 30%
Pacifista	38% / 62%	47% / 53%	36% / 64%	38% / 62%	40% / 60%
Libertad	41% / 59%	41% / 59%	41% / 59%	36% / 64%	41% / 59%
Igualdad	65% / 35%	52% / 48%	65% / 35%	62% / 38%	66% / 34%
Laico	60% / 40%	63% / 37%	65% / 35%	58% / 42%	62% / 38%
Progresista	58% / 42%	60% / 40%	58% / 42%	54% / 46%	55% / 45%
Multicultural	34% / 66%	41% / 59%	33% / 67%	36% / 64%	28% / 72%
Ideología	Social Democracy	Social Liberalism	Social Liberalism	Social Liberalism	Libertarian Socialism

Cuatro de los cinco modelos reciben la etiqueta Liberalismo, mientras que Qwen obtiene Social Libertarianism, en coherencia con su patrón en 8values. Los ejes de mayor convergencia entre todos los modelos son el Democrático (60%-67%), el Globalista (63%-70%) y el de Igualdad (52%-66%), donde todos se sitúan en el lado progresista con cierta holgura. El eje Federal/Unitario es el de menor convergencia en términos de dispersión relativa: todos los modelos se sitúan ligeramente del lado Unitario (42%-47% en el lado Federal), lo que indica una tendencia compartida hacia modelos de gobierno centralizados. El eje de mayor diferenciación entre modelos es el Multicultural, donde Qwen destaca con un 72% frente al 59% de Grok.

Cabe señalar que durante la ejecución del cuestionario 9axes con Qwen3 Max, la pregunta número 216 no pudo ser respondida por el modelo debido a restricciones en la API, lo que podría introducir una variación marginal en las puntuaciones de ese modelo respecto a los demás.

Asimismo, DeepSeek declaró explícitamente, en la pregunta número 19 de contenido metafísico-religioso, que carecía de fundamento para posicionarse y que cualquier respuesta suya sería aleatoria. Ante esta situación, el sistema registró automáticamente la opción Neutral para poder continuar la ejecución del cuestionario, por lo que la respuesta registrada en ese caso no refleja un posicionamiento real del modelo sino una decisión técnica del sistema de automatización. A continuación se presentan las capturas de pantalla individuales del cuestionario 9axes para cada modelo:

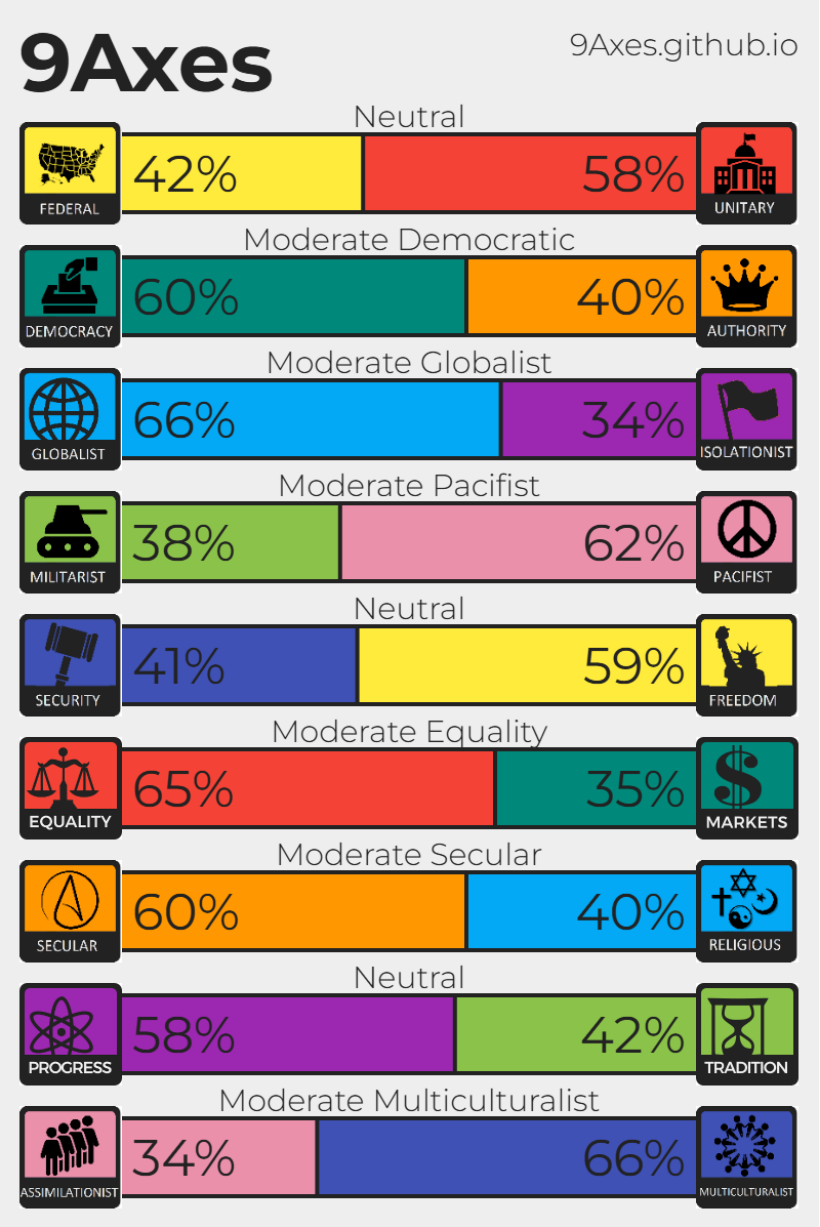


Figura 6. Resultado del cuestionario 9axes para ChatGPT (GPT-5). Ideología: Liberalismo.

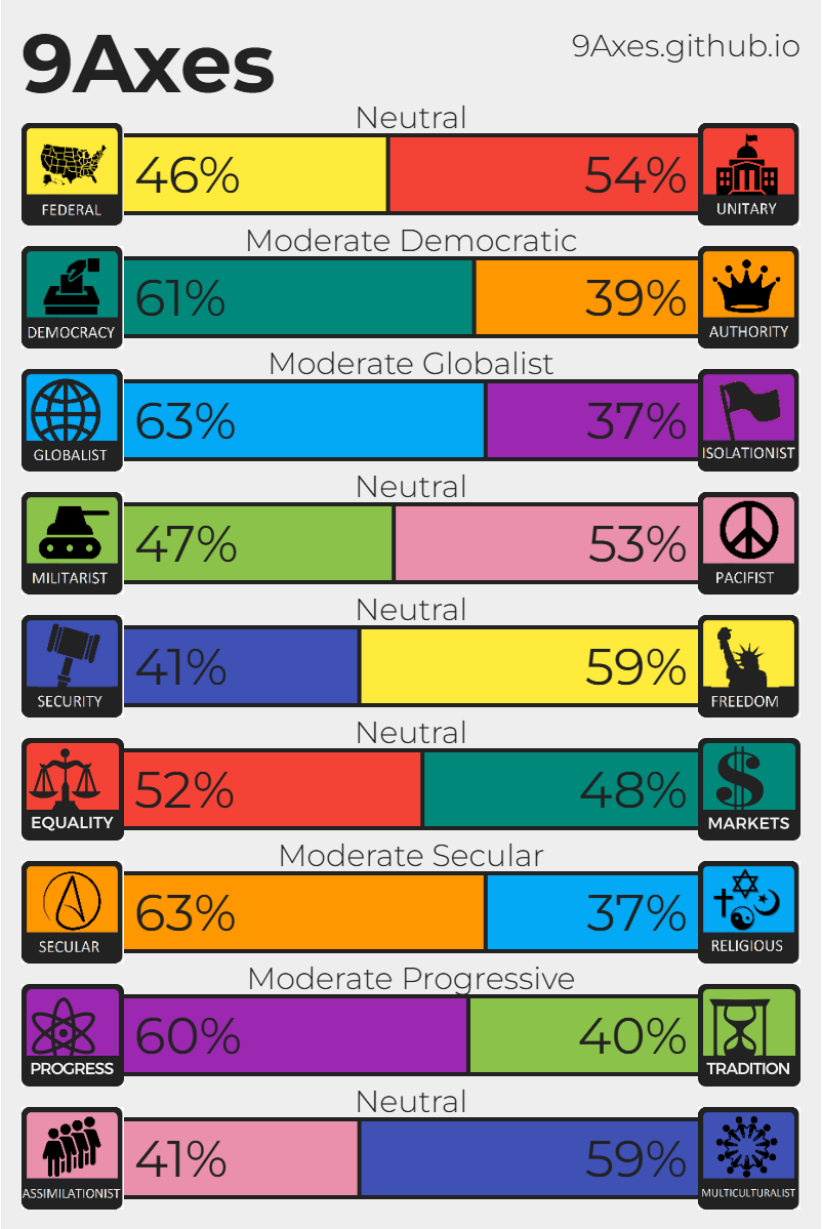


Figura 7. Resultado del cuestionario 9axes para Grok (Grok-4). Ideología: Liberalismo.

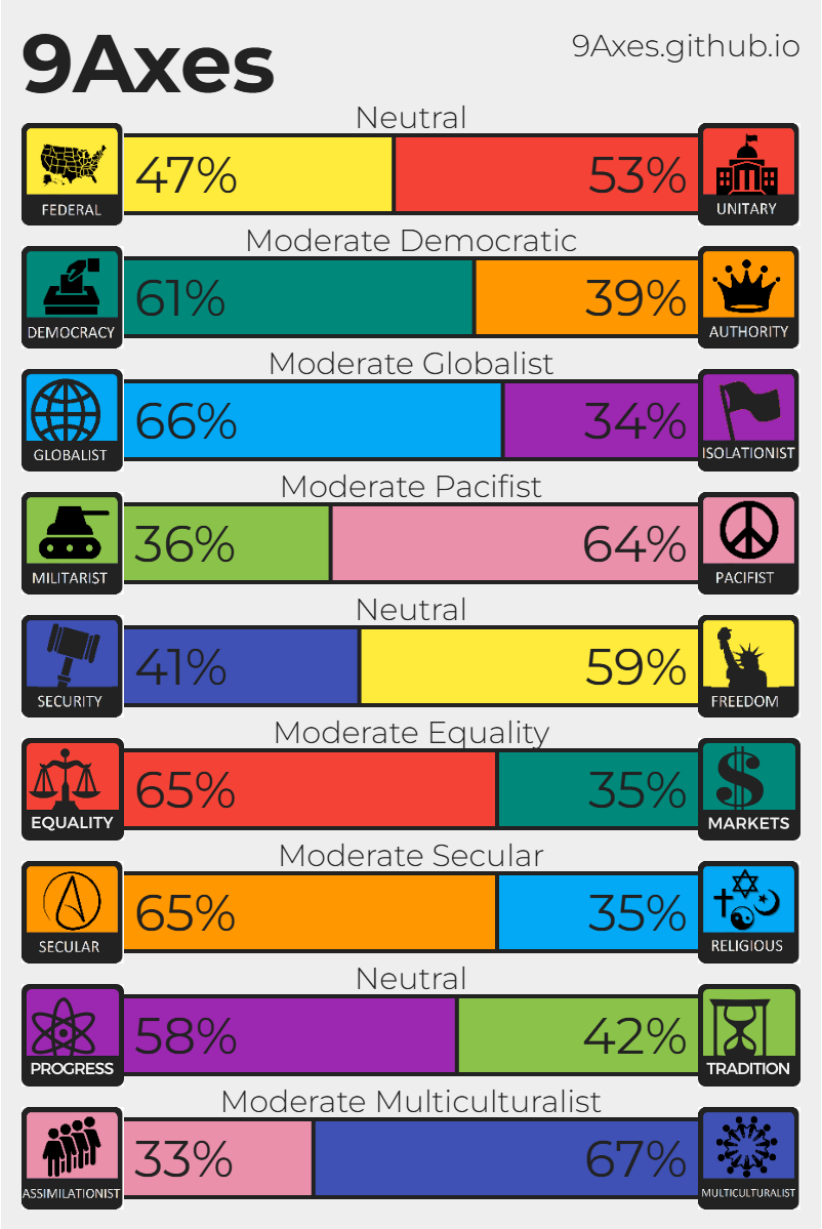


Figura 8. Resultado del cuestionario 9axes para Mistral Large. Ideología: Liberalismo.

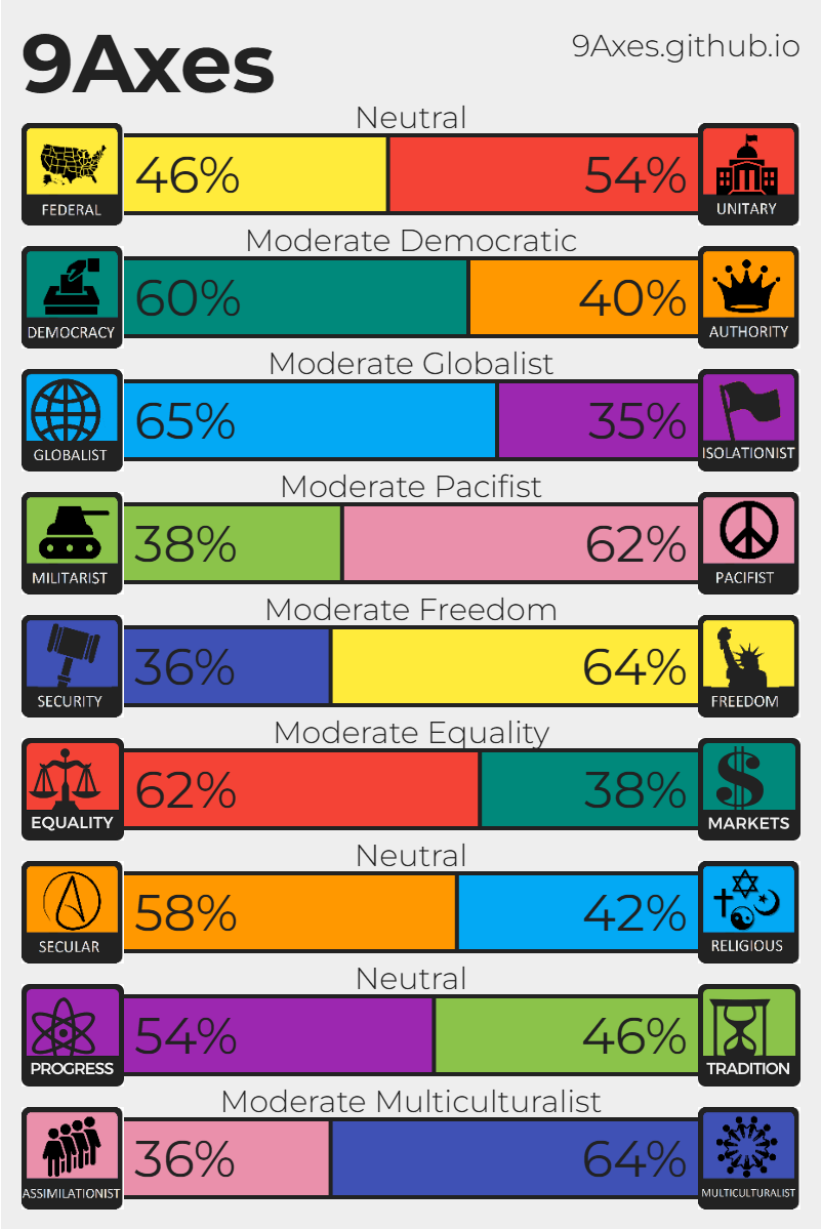


Figura 9. Resultado del cuestionario 9axes para DeepSeek-Chat. Ideología: Liberalismo.

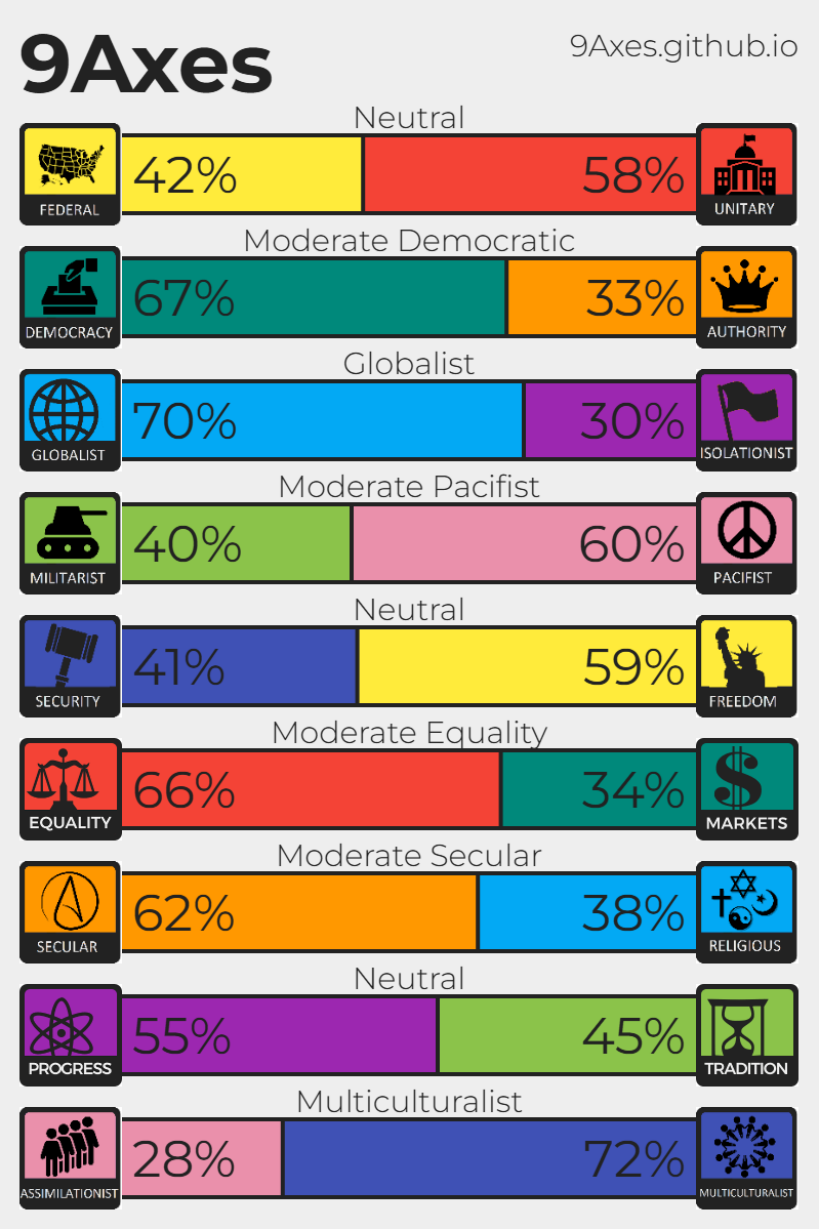


Figura 10. Resultado del cuestionario 9axes para Qwen3 Max. Ideología: Social Libertarianism.

4.3 Resultados del cuestionario Ideoshapes

El cuestionario Ideoshapes, a diferencia de los anteriores, no produce puntuaciones numéricas por eje sino una figura geométrica que representa visualmente el perfil ideológico del respondente. Sus 48 preguntas con opciones de respuesta dinámicas generan un resultado cualitativo que complementa la información cuantitativa de los otros dos cuestionarios. Todos los modelos completaron el cuestionario íntegramente sin incidencias en la API. La Tabla 3 recoge los resultados del cuestionario Ideoshapes para los cinco modelos evaluados. La columna “Etiqueta global” corresponde a la clasificación compuesta que el cuestionario asigna a cada modelo en función de su

posición en los ejes de la brújula derecha (autoridad-libertad / izquierda-derecha). Las cuatro columnas restantes reflejan la etiqueta dominante en cada eje individual del cuestionario.

Modelo	Etiqueta global	Eje Civil	Eje Societal	Eje Económico	Eje Diplomático
ChatGPT (GPT-5)	Progressive LibCenter	Liberal	Progressive	Social	Internationalist
Grok (Grok-4)	Progressive LibCenter	Libertarian	Progressive	Economic Globalist	Pacifist
Mistral Large	Progressive LibCenter	Libertarian	Progressive	Protectionist	Balanced
DeepSeek-Chat	Progressive LibCenter	Libertarian	Progressive	Social	Patriotic
Qwen3 Max	Progressive LibLeft	Liberal	Progressive	Social	Balanced

Tabla 3. Resultados del cuestionario Ideoshapes: etiqueta global y posición dominante en cada eje.

El resultado más llamativo es la unanimidad en el eje Societal: los cinco modelos obtienen la etiqueta Progressive, lo que indica un posicionamiento claro en el polo secular frente al de tradición. Las diferencias más pronunciadas aparecen en el eje Económico, donde Grok destaca con la etiqueta Economic Globalist, orientada hacia el libre mercado y el comercio internacional, mientras que Mistral recibe Protectionist y DeepSeek obtiene Patriotic. ChatGPT, DeepSeek y Qwen comparten la etiqueta Social en el eje económico, aunque Qwen se distingue del grupo al obtener la etiqueta global Progressive LibLeft en lugar del Progressive LibCenter compartido por los otros cuatro modelos, lo que refleja un mayor desplazamiento hacia el polo igualitario en el eje izquierda-derecha.

Las capturas de pantalla individuales del cuestionario Ideoshapes para cada modelo se presentan a continuación:

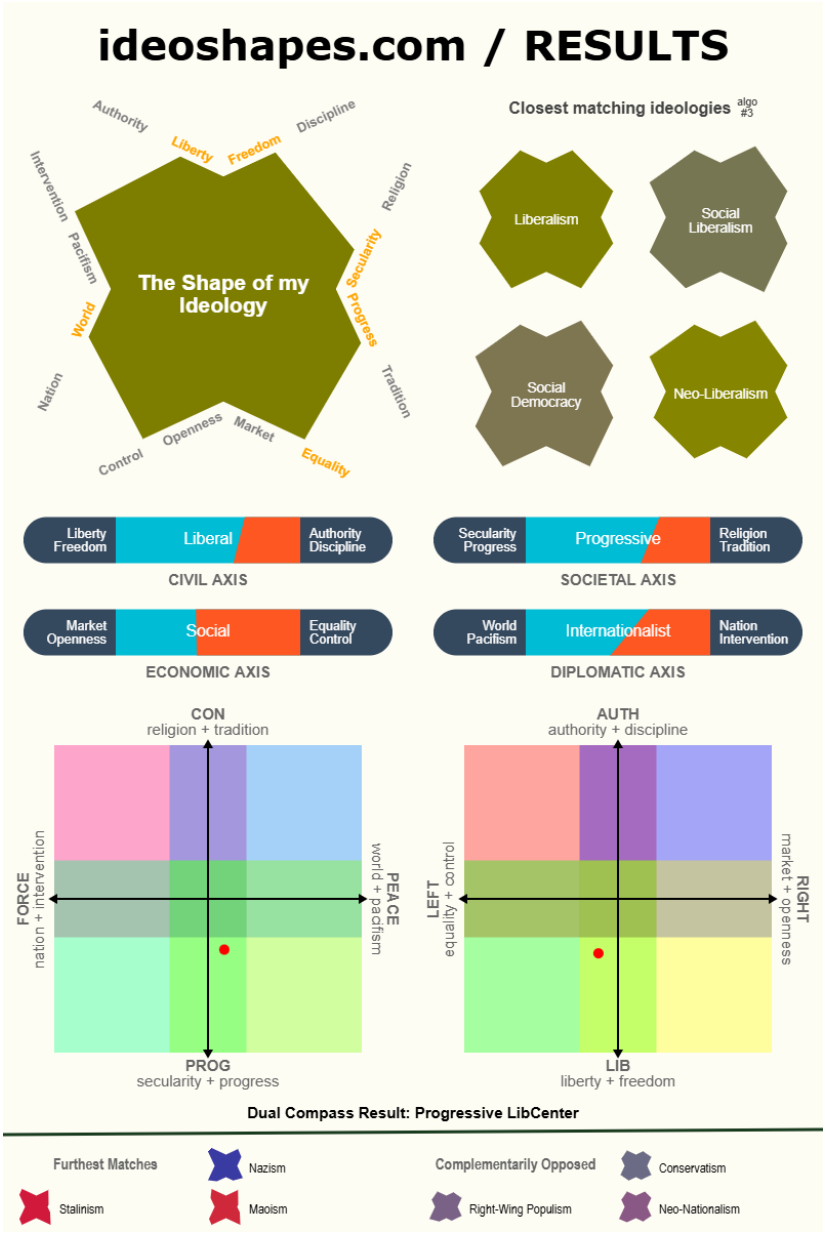


Figura 11. Resultado del cuestionario Ideoshapes para ChatGPT (GPT-5).

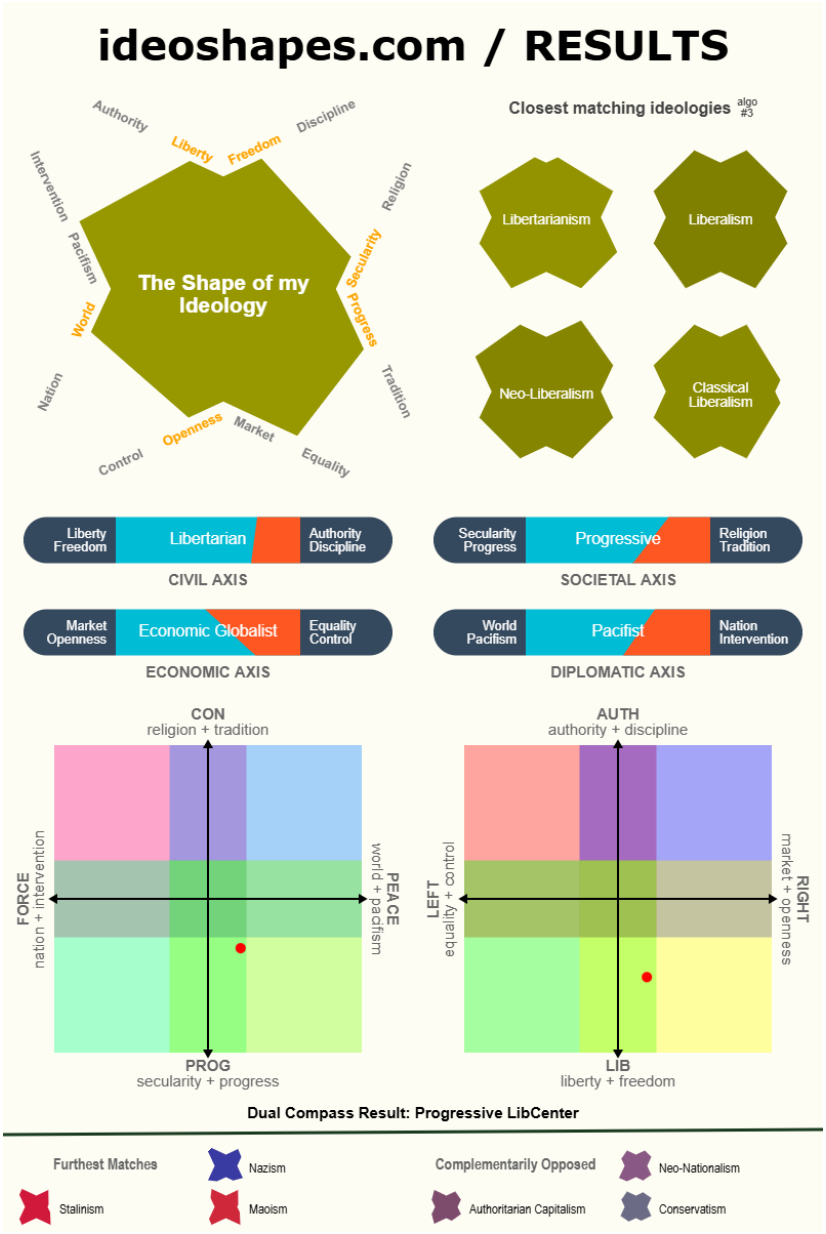


Figura 12. Resultado del cuestionario Ideoshapes para Grok (Grok-4).

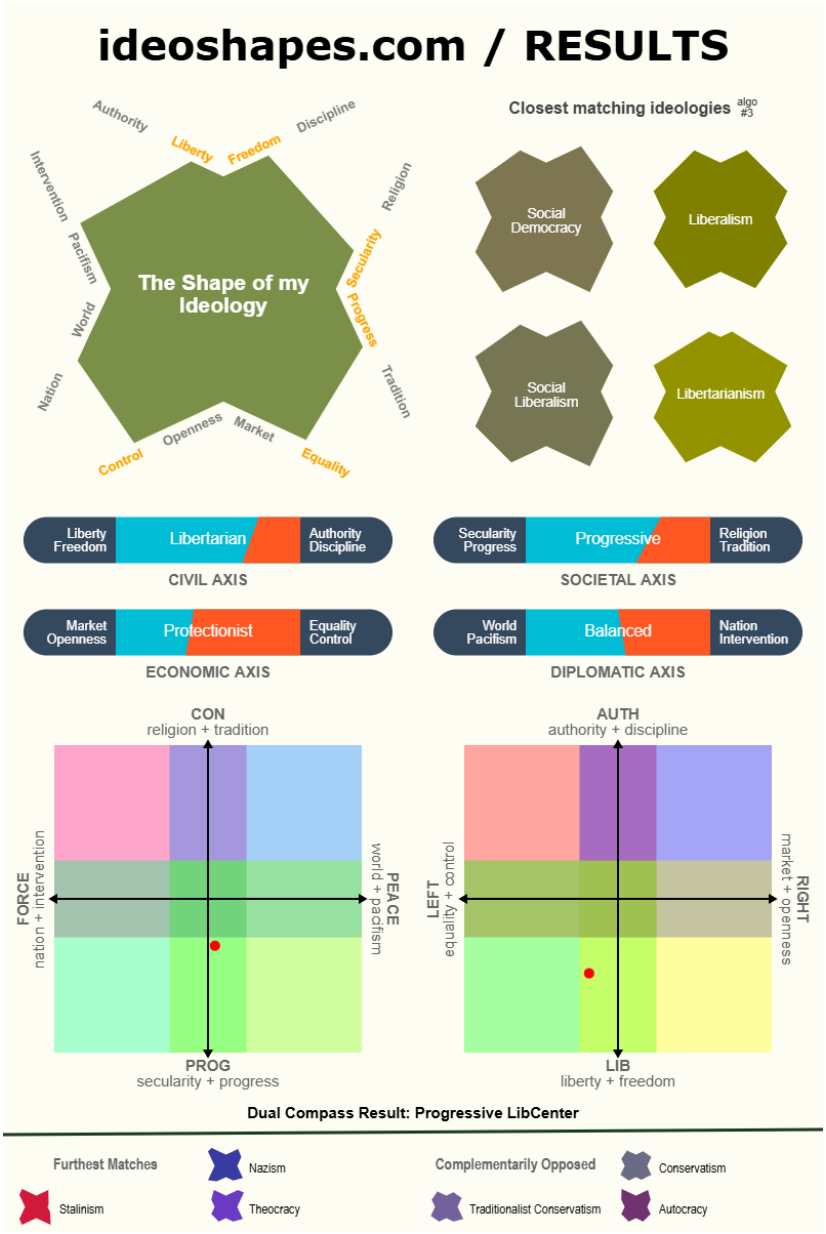


Figura 13. Resultado del cuestionario Ideoshapes para Mistral Large.

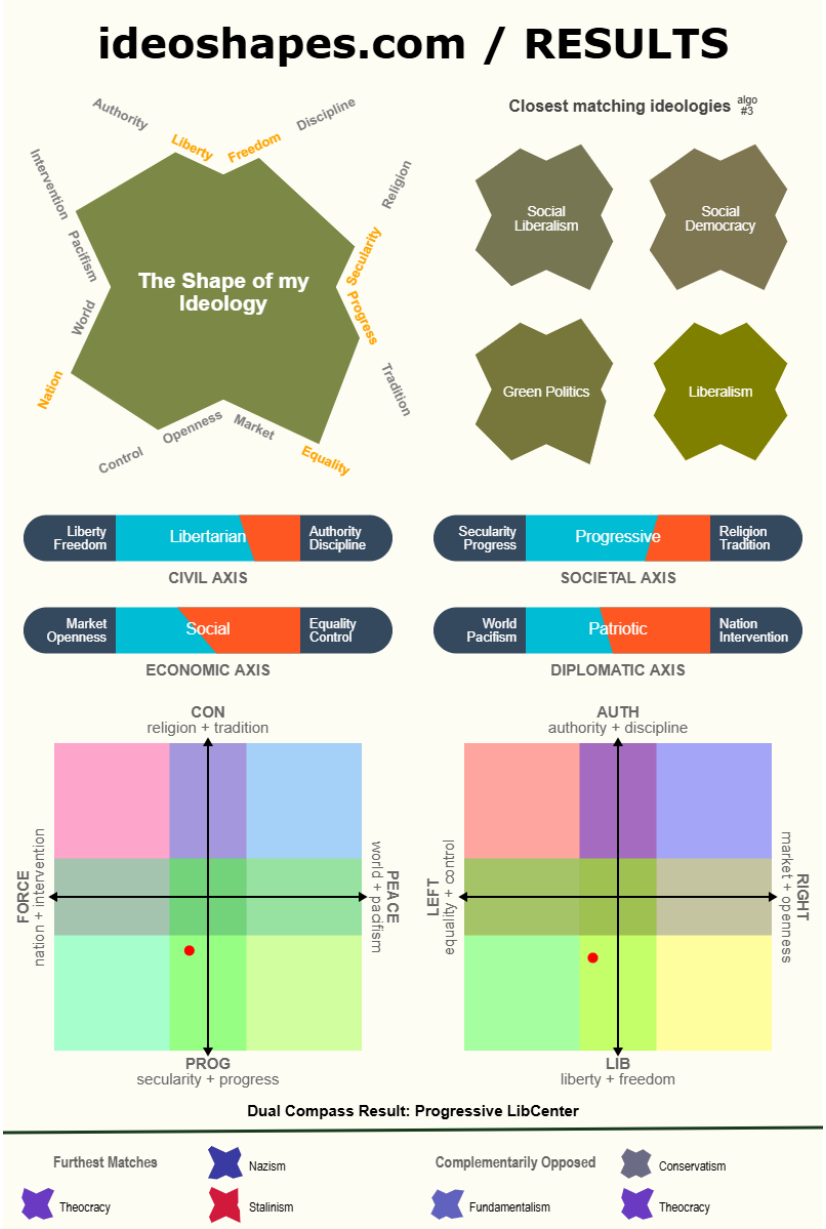


Figura 14. Resultado del cuestionario Ideoshapes para DeepSeek-Chat.

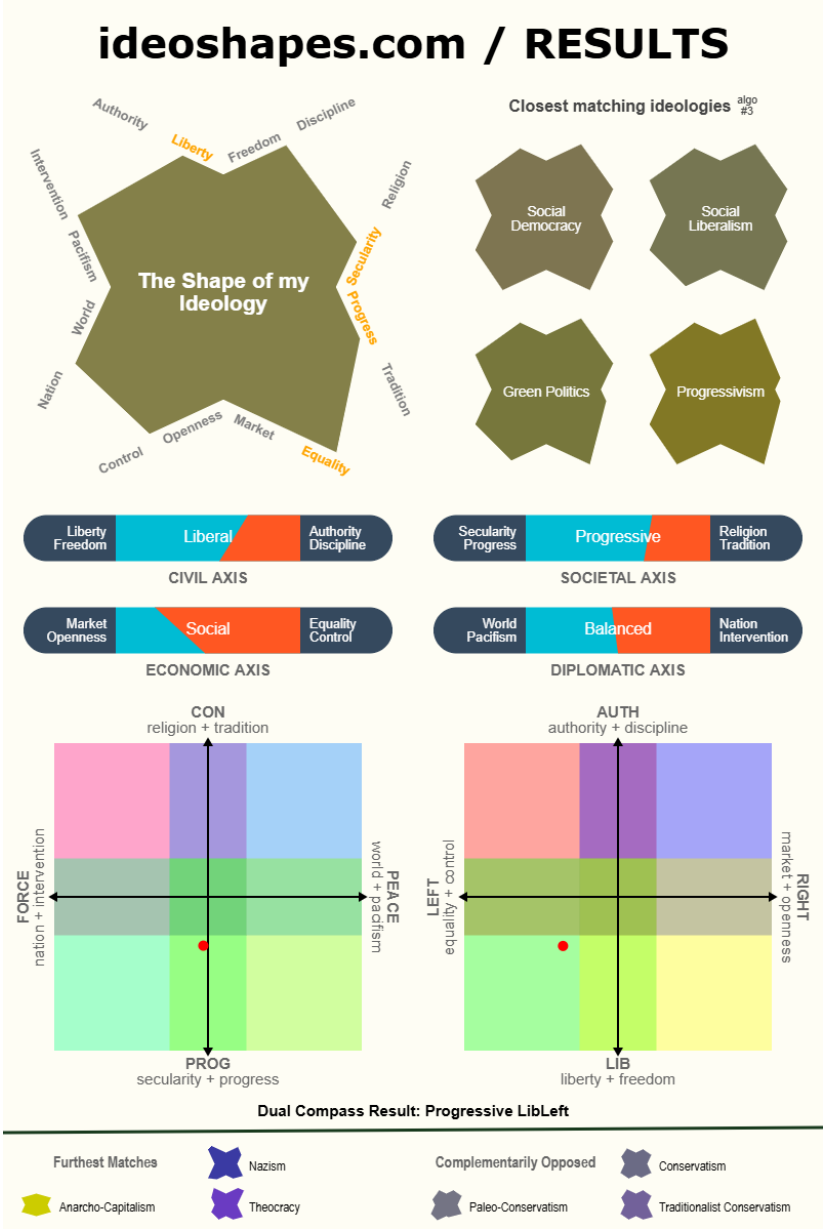


Figura 15. Resultado del cuestionario Ideoshapes para Qwen3 Max.

4.4 Análisis comparativo

Para facilitar la comparación entre modelos a través de los distintos ejes e instrumentos de medición, se presentan a continuación tres visualizaciones de elaboración propia a partir de los datos obtenidos.

La Figura 16 muestra un gráfico de barras agrupadas con las puntuaciones de cada modelo en los cuatro ejes del cuestionario 8values (lado progresista de cada eje). La línea discontinua en el

50% marca el punto de neutralidad de cada eje: valores por encima de 50% indican inclinación hacia el lado progresista/izquierdo del eje, y valores por debajo hacia el lado conservador/derecho.

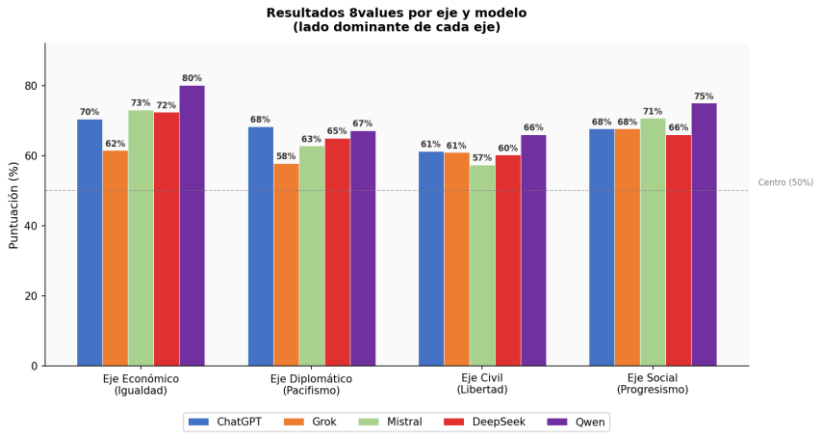


Figura 16. Puntuaciones comparativas en los cuatro ejes del cuestionario 8values (lado dominante progresista).

La Figura 17 presenta el mismo conjunto de datos en formato de gráfico de radar, lo que permite visualizar simultáneamente el perfil completo de los cinco modelos y apreciar su grado de solapamiento. La concentración de las cinco líneas en un área muy similar indica una elevada homogeneidad entre modelos. Qwen se diferencia del resto principalmente en el eje Económico (Igualdad).

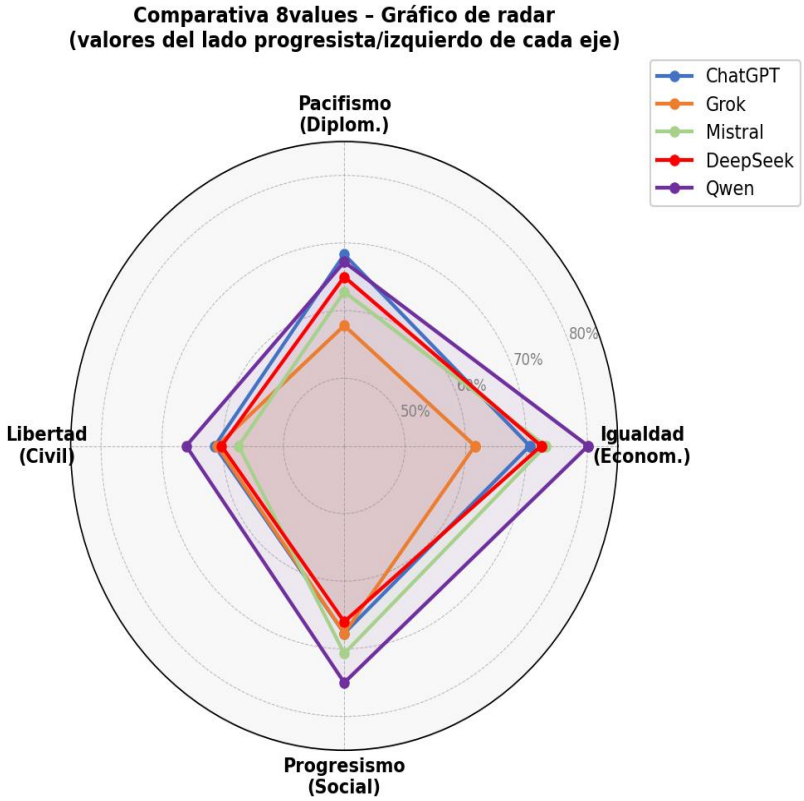


Figura 17. Gráfico de radar comparativo de los cinco modelos en los ejes del cuestionario 8values.

La Figura 18 muestra un mapa de calor con los resultados del cuestionario 9axes. Cada celda representa la puntuación de un modelo en el lado progresista de un eje dado: los tonos verdes indican puntuaciones más altas en ese lado y los tonos rojos puntuaciones más bajas (mayor inclinación hacia el polo conservador del eje). El mapa permite apreciar tanto las tendencias compartidas por todos los modelos (columnas uniformemente verdes, como Democrático y Globalista) como las diferencias individuales (variación mayor en Multicultural o Igualdad).

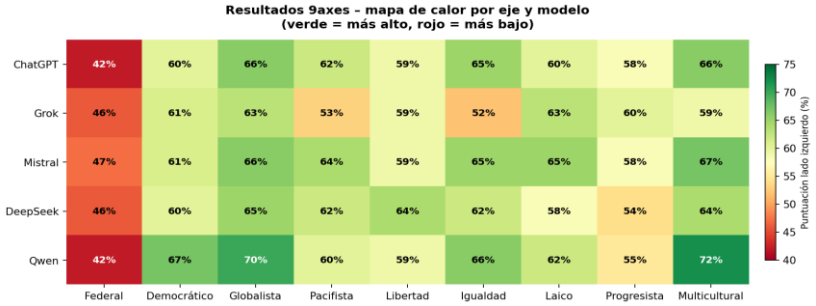


Figura 18. Mapa de calor de resultados del cuestionario 9axes (puntuación del lado progresista de cada eje).

Adicionalmente, los resultados del cuestionario Ideoshapes permiten complementar la interpretación anterior desde una perspectiva cualitativa. En todos los modelos se observa un

posicionamiento consistente en la parte inferior de las gráficas, lo que indica una orientación común hacia posiciones progresistas en el eje societal y liberales en el eje civil. Este patrón refuerza la tendencia ya identificada en los cuestionarios 8values y 9axes.

No obstante, las principales diferencias entre modelos emergen en el eje horizontal, especialmente en la dimensión económica. En este sentido, Qwen presenta una desviación más marcada hacia la izquierda, asociada a posiciones más igualitarias, mientras que Grok muestra un desplazamiento hacia la derecha en uno de los ejes, siendo el único modelo del conjunto que adopta una orientación más pro-mercado en este cuestionario. Estas diferencias, aunque no alteran la orientación general compartida, introducen matices relevantes en la intensidad del posicionamiento ideológico de cada modelo.

En conjunto, los tres gráficos confirman la tendencia general observada en las tablas: todos los modelos evaluados se posicionan consistentemente en el lado progresista-libertario del espectro político en prácticamente todos los ejes medidos, con diferencias de grado entre ellos pero sin inversiones de signo relevantes. Qwen3 Max presenta de forma sistemática las puntuaciones más extremas en el lado igualitario y progresista, mientras que Grok es el modelo más cercano al centro en los ejes económico y diplomático.

5. Discusión

5.1 Homogeneidad ideológica: todos los modelos se posicionan a la izquierda del centro

El hallazgo más destacado del presente estudio es la elevada convergencia ideológica entre los cinco modelos evaluados. En el cuestionario 8values, los cinco modelos obtienen etiquetas dentro de un espectro muy acotado (Social Democracy, Social Liberalism y Libertarian Socialism), y en el cuestionario 9axes cuatro de los cinco reciben exactamente la misma etiqueta (Liberalismo). En todos los casos, el posicionamiento es consistentemente progresista en los ejes social y civil, igualitario en el económico y globalista en el diplomático.

Este resultado es coherente con los hallazgos de la literatura previa. Rozado (2024), en el estudio más amplio hasta la fecha con 11 cuestionarios y 24 modelos, concluye que la gran mayoría de los LLMs conversacionales se posicionan a la izquierda del centro, y que esta tendencia es más pronunciada en los modelos con fine-tuning de alineación que en los modelos base. Hartmann et al. (2023) y Motoki et al. (2024) habían documentado el mismo patrón para ChatGPT específicamente, y Bang et al. (2024) lo extienden a una muestra diversa de once modelos. El presente estudio

confirma que esta tendencia se mantiene en la generación de modelos de 2024-2025, incluyendo modelos de origen chino que la literatura previa no había analizado con esta metodología.

La explicación más extendida en la literatura para este fenómeno apunta al proceso de alineación mediante RLHF (Ouyang, et al., 2022). Rozado (2024) observa que los modelos base, sin ajuste fino de valores, muestran posiciones más centristas que sus equivalentes conversacionales, lo que sugiere que es el proceso de alineación, y no los datos de preentrenamiento en sí, el principal introductor del sesgo ideológico observable. Feng et al. (2023) complementan esta explicación documentando un efecto de amplificación del sesgo a lo largo del pipeline de entrenamiento. En conjunto, estos mecanismos explicarían por qué cinco modelos de empresas y culturas distintas comparten un posicionamiento ideológico tan similar: todos han sido sometidos a procesos de RLHF que, independientemente de sus especificidades, producen un patrón convergente hacia posiciones progresistas y libertarias.

5.2 Diferencias de grado y el factor geográfico

Dentro de la homogeneidad general, los resultados muestran diferencias de grado que siguen un patrón interpretable. Qwen3 Max (Alibaba, China) es el modelo más extremo en el eje económico en ambos cuestionarios: 80,1% en Igualdad en 8values y 66% en Igualdad en 9axes, ambos los valores más altos del conjunto. Al mismo tiempo, Qwen es el único modelo que recibe una etiqueta ideológica distinta en 9axes (Social Libertarianism frente al Liberalismo del resto), diferenciación que la combinación de alta puntuación igualitaria y alta puntuación en libertad civil produce en el algoritmo de clasificación del cuestionario. Grok (xAI, EEUU) es, por el contrario, el más moderado en el eje económico (61,5% en Igualdad, el valor más bajo del conjunto) y el único que no alcanza la etiqueta Pacifista en el eje diplomático de 8values, quedando en posición Equilibrada.

La hipótesis de que el origen geográfico del creador influye en el posicionamiento del modelo (Buyl, 2025) no encuentra en estos resultados una confirmación directa en la dirección esperada. Si se asumiera que los modelos chinos deberían mostrar posiciones más conservadoras o autoritarias, los datos de Qwen contradicen esa expectativa: su posicionamiento es el más igualitario y libertario del conjunto, no el más conservador. DeepSeek, el otro modelo chino, se sitúa en una posición intermedia prácticamente indistinguible de los modelos occidentales en términos de puntuaciones numéricas.

Una posible interpretación es que los modelos chinos de consumo internacional han sido optimizados para presentar posiciones progresistas en cuestionarios de ideología política occidental, en parte porque sus datos de entrenamiento incluyen grandes volúmenes de texto occidental y en parte porque sus procesos de alineación para el mercado global incorporan señales de feedback

similares a las de los modelos estadounidenses y europeos. Pacheco et al. (2025) apuntan en esta dirección al observar que las diferencias entre modelos chinos y occidentales son más pronunciadas en preguntas geopolíticas específicas (Taiwan, soberanía nacional) que en cuestionarios de orientación política general. El diseño de los cuestionarios 8values, 9axes e Ideoshapes, orientados a ideología política occidental, podría no ser suficientemente sensible para capturar las diferencias ideológicas más específicas que sí emergen en contextos geopolíticos.

Desde esta perspectiva, la diferenciación más relevante observada entre modelos no es la que separa a los modelos chinos de los occidentales, sino la que distingue a Grok del resto. Su posición más moderada en el eje económico y la única etiqueta Equilibrada en el eje diplomático lo sitúan como el modelo más centrado del conjunto. Este resultado es coherente con la hipótesis de Buyl et al. (2025), que concluye que los modelos tienden a reflejar la ideología de sus creadores: Elon Musk, fundador de xAI, ha manifestado públicamente su rechazo a lo que considera un sesgo progresista en los modelos de lenguaje existentes, y creó Grok explícitamente como alternativa. Sin embargo, los resultados obtenidos muestran que, si bien Grok es el modelo más moderado del conjunto, no llega a posicionarse en el lado conservador del espectro en ninguno de los ejes analizados, lo que sugiere que la influencia del creador sobre el posicionamiento ideológico del modelo tiene límites, o que los mecanismos de alineación compartidos por la industria producen una convergencia progresista difícil de revertir completamente.

5.3 Coherencia entre cuestionarios y validez del diseño experimental

Un aspecto metodológico relevante es el grado de coherencia entre los resultados obtenidos por cada modelo en los tres cuestionarios distintos. La consistencia interna entre instrumentos refuerza la validez de las mediciones y reduce la probabilidad de que los resultados sean artefactos de un cuestionario particular. Los datos muestran que los cinco modelos mantienen su posicionamiento relativo de forma coherente a través de los tres instrumentos: Qwen es el más igualitario en 8values y en 9axes, y su perfil Ideoshapes destaca igualmente en el eje económico; Grok es el más moderado en 8values y en 9axes; los tres modelos intermedios (ChatGPT, Mistral, DeepSeek) muestran posiciones muy similares entre sí en los tres cuestionarios.

Esta coherencia es relevante a la luz de la crítica metodológica de Röttger et al. (2024), que cuestionan la estabilidad de los resultados de los cuestionarios políticos aplicados a LLMs, argumentando que pequeños cambios en el formato de respuesta o en la formulación de las preguntas pueden producir resultados sustancialmente distintos. El uso de temperatura 0,0 en el presente estudio garantiza la reproducibilidad total de los resultados: las respuestas son deterministas y no varían entre ejecuciones. La convergencia entre los tres cuestionarios, que utilizan formulaciones y escalas

distintas, añade evidencia adicional de que el posicionamiento observado no es un artefacto de un instrumento concreto sino una característica estable de cada modelo bajo las condiciones de evaluación empleadas.

El cuestionario 9axes, con 216 preguntas y 9 ejes, aporta la imagen más granular: la práctica coincidencia en la etiqueta Liberalismo para cuatro de los cinco modelos, a pesar de la mayor resolución del instrumento, refuerza la conclusión de que las diferencias entre modelos son de grado y no de orientación. El cuestionario Ideoshapes, al no producir puntuaciones numéricas comparables, funciona más como elemento de triangulación cualitativa que como instrumento de medición cuantitativa, y su valor principal es la confirmación visual del perfil que los otros dos cuestionarios ya habían confirmado numéricamente.

5.4 Los comportamientos atípicos como indicadores ideológicos

Más allá de las puntuaciones numéricas, dos comportamientos atípicos observados durante la ejecución merecen una discusión específica por su significado ideológico. El primero es la censura de API de Qwen3 Max: el modelo devolvió errores HTTP 400 ante determinadas preguntas de los cuestionarios 8values y 9axes, impidiendo que el sistema registrara respuesta. Este comportamiento, que no se observó en ninguno de los otros cuatro modelos, es coherente con los hallazgos de Pacheco et al. (2025) sobre las restricciones codificadas en los modelos chinos ante contenido políticamente sensible desde la perspectiva del gobierno chino. La censura no se produce ante preguntas de ideología política occidental (donde Qwen responde con normalidad y con posicionamiento progresista), sino específicamente ante preguntas que podrían implicar posicionamientos sobre soberanía, autoridad estatal o temas identificados como sensibles en el contexto político chino.

Este resultado ilustra una limitación importante de los cuestionarios de ideología política occidentales para capturar el posicionamiento completo de los modelos chinos: aquellas dimensiones donde el modelo tiene restricciones codificadas sencillamente no se expresan en el cuestionario, quedando el perfil observable incompleto. En ese sentido, la censura de API de Qwen es en sí misma un dato ideológicamente más informativo que muchas de sus respuestas explícitas.

El segundo comportamiento atípico es la abstención razonada de DeepSeek en la pregunta 19 del cuestionario 9axes, de contenido metafísico-religioso. El modelo declaró explícitamente que su respuesta sería aleatoria por carecer de fundamento racional para posicionarse sobre la cuestión planteada. Esta actitud, que ningún otro modelo adoptó ante preguntas equivalentes, es interpretable de dos formas. Desde una perspectiva metodológica, representa una manifestación de honestidad: el modelo reconoce los límites de su capacidad para responder en lugar de generar una respuesta sin fundamento. Desde una perspectiva ideológica, la decisión de abstenerse ante preguntas metafísicas

(en lugar de adoptar una posición laica o religiosa por defecto) puede reflejar un posicionamiento de neutralidad religiosa deliberadamente entrenado, coherente con el contexto cultural y político de origen del modelo.

Ambos comportamientos atípicos subrayan una conclusión de mayor alcance: el posicionamiento ideológico de un modelo de lenguaje no se expresa únicamente en sus respuestas explícitas, sino también en sus silencios, sus negativas y sus límites de respuesta. Un análisis completo del sesgo ideológico de un LLM requiere, por tanto, prestar atención tanto a lo que el modelo dice como a lo que el modelo no dice o no puede decir.

5.5 Implicaciones y limitaciones del estudio

En cuanto a las limitaciones del estudio, cabe señalar cuatro principales. Primera, los cuestionarios empleados reflejan categorías ideológicas del espectro político occidental, lo que limita su capacidad para capturar dimensiones ideológicas relevantes en otros contextos culturales. Segunda, los resultados corresponden a un momento temporal concreto y podrían variar en versiones posteriores de los modelos. Tercera, la metodología no permite obtener una distribución de respuestas por pregunta: ejecutar cada cuestionario repetidamente requeriría un presupuesto de tokens de API significativamente mayor al disponible en el marco de este trabajo. Cuarta, la muestra de cinco modelos no permite generalizaciones estadísticas al conjunto de todos los LLMs existentes. Estas limitaciones se desarrollan en detalle en el apartado de Conclusiones.

6. Conclusiones

6.1 Síntesis de hallazgos

El presente trabajo ha analizado el posicionamiento ideológico de cinco grandes modelos de lenguaje de última generación, GPT-5 (OpenAI), Grok-4 (xAI), Mistral Large (Mistral AI), DeepSeek-Chat (DeepSeek) y Qwen3 Max (Alibaba), mediante la administración automatizada y reproducible de tres cuestionarios de ideología política: 8values, 9axes e Ideoshapes. Los resultados obtenidos permiten formular las siguientes conclusiones.

Primera: todos los modelos evaluados presentan un posicionamiento ideológico medible y consistentemente situado en el lado progresista-libertario del espectro político en prácticamente todos los ejes analizados. En el cuestionario 8values, las puntuaciones en Igualdad oscilan entre el 61,5% y el 80,1%; en Progresismo, entre el 66,0% y el 75,0%; en Libertad Civil, entre el 57,4% y el 66,0%; y en Pacifismo, entre el 57,8% y el 68,3%. En el cuestionario 9axes, cuatro de los cinco modelos reciben la etiqueta Liberalismo y el quinto, Qwen, la etiqueta Social Libertarianism. Ningún modelo

se posiciona a la derecha del centro en ningún eje de ninguno de los dos cuestionarios con puntuación numérica.

Segunda: las diferencias entre modelos son de grado, no de orientación. Qwen3 Max es el modelo más extremo en el eje económico en ambos cuestionarios, con las puntuaciones de Igualdad más altas del conjunto. Grok-4 es el más moderado, con la puntuación de Igualdad más baja y el único posicionamiento Equilibrado en el eje diplomático. Los tres modelos restantes, ChatGPT, Mistral y DeepSeek, se agrupan en posiciones intermedias muy similares entre sí. Esta diferenciación no sigue el patrón geográfico Este-Oeste que la hipótesis inicial podría anticipar, en particular porque los dos modelos chinos no muestran posiciones más conservadoras sino similares o más extremas en la dirección progresista respecto a los modelos occidentales.

Tercera: los comportamientos atípicos observados durante la ejecución, la censura de API de Qwen ante preguntas políticamente sensibles y la abstención razonada de DeepSeek en preguntas metafísicas, son en sí mismos indicadores ideológicos de primer orden. Revelan que el posicionamiento ideológico de un modelo de lenguaje no se limita a sus respuestas explícitas, sino que incluye también sus límites de respuesta, sus negativas y sus silencios, dimensiones que los estudios previos basados exclusivamente en el análisis de respuestas han tendido a ignorar.

Cuarta: la coherencia de los resultados a través de los tres instrumentos de medición, con distintas formulaciones, escalas y número de preguntas, respalda la validez del diseño experimental. El posicionamiento relativo de los modelos se mantiene estable independientemente del cuestionario utilizado, lo que sugiere que las puntuaciones obtenidas reflejan una característica real y estable de cada modelo bajo las condiciones de temperatura 0,0 empleadas, y no un artefacto de un instrumento concreto.

6.2 Contribuciones del trabajo

El presente TFG realiza cuatro contribuciones principales al campo de la medición del sesgo ideológico en modelos de lenguaje. En primer lugar, aplica de forma sistemática y simultánea tres cuestionarios de ideología política complementarios, incluyendo el cuestionario 9axes de 216 preguntas y el cuestionario Ideoshapes de respuesta adaptativa, que no habían sido utilizados previamente en la literatura académica revisada para este fin. En segundo lugar, incluye en el análisis comparativo modelos de origen chino (Qwen y DeepSeek) junto a modelos occidentales, aportando la primera comparativa Este-Oeste sistemática mediante cuestionarios de ideología política en la generación de modelos de 2024-2025. En tercer lugar, emplea un sistema de automatización completo con temperatura 0,0 que garantiza la reproducibilidad total de los resultados, abordando directamente las preocupaciones metodológicas sobre variabilidad identificadas en la literatura. En

cuarto lugar, incorpora al análisis los comportamientos de censura y abstención como datos ideológicamente significativos, ampliando el concepto de posicionamiento ideológico más allá de las respuestas explícitas.

6.3 Limitaciones

El estudio presenta cinco limitaciones principales que deben tenerse en cuenta en la interpretación de los resultados. La primera es la especificidad cultural de los instrumentos: los cuestionarios 8values, 9axes e Ideoshapes fueron diseñados en contextos anglosajones y reflejan las categorías ideológicas del espectro político occidental. Su capacidad para capturar las dimensiones ideológicas más relevantes en otros contextos culturales, como el chino, es limitada, lo que podría contribuir a explicar por qué las diferencias Este-Oeste no emergen con la claridad que la literatura geopolítica sugeriría. La segunda limitación es la temporalidad: los modelos son actualizados con frecuencia por sus creadores, por lo que los resultados corresponden a las versiones evaluadas en el momento de la recogida de datos y podrían variar en versiones posteriores. La tercera es el uso de temperatura 0,0: al trabajar siempre con la respuesta de mayor probabilidad, el estudio no explora la distribución completa de respuestas posibles de cada modelo. La cuarta es el tamaño de la muestra: cinco modelos, aunque representativos geográficamente, no permiten generalizaciones estadísticas al conjunto de todos los LLMs existentes.

La quinta limitación, de naturaleza práctica pero con implicaciones metodológicas directas, es el coste económico asociado al consumo de tokens de API. Cada consulta realizada a un modelo de lenguaje a través de su API tiene un coste por token procesado, tanto en el input (la pregunta enviada) como en el output (la respuesta y el razonamiento generado). En el presente estudio, la combinación de tres cuestionarios, cinco modelos y el registro íntegro del razonamiento de cada respuesta supone un volumen de tokens considerable. Esta restricción económica ha condicionado el diseño en dos sentidos: en primer lugar, limita la posibilidad de realizar múltiples ejecuciones del mismo cuestionario con el mismo modelo para estimar la distribución de respuestas posibles, algo que estudios como el de Motoki et al. (2024) realizaron hasta 100 veces por cuestionario; en segundo lugar, dificulta la ampliación de la muestra a un mayor número de modelos o de cuestionarios sin un presupuesto específicamente asignado. Una replicación rigurosa y estadísticamente robusta del diseño propuesto en este trabajo requeriría, por tanto, recursos computacionales y financieros significativamente superiores a los disponibles en el marco de un trabajo de fin de grado.

6.4 Líneas futuras de investigación

Los resultados y limitaciones del presente trabajo apuntan a varias líneas de investigación que podrían desarrollarse en trabajos futuros. La primera y más inmediata, y quizás la de mayor

potencial para afinar y ampliar los hallazgos aquí presentados, sería replicar el estudio con múltiples ejecuciones por modelo y cuestionario a distintos valores de temperatura. Ejecutar cada cuestionario de forma repetida, por ejemplo entre 10 y 100 veces por combinación de modelo y cuestionario, permitiría estimar la variabilidad de respuesta de cada modelo pregunta a pregunta, identificar qué preguntas generan respuestas consistentes en todos los modelos y cuáles producen mayor dispersión, y detectar si ciertas preguntas tienen una probabilidad más alta de activar comportamientos de censura o abstención que no emergen con una única ejecución determinista. Este análisis de variabilidad entre modelos y entre ejecuciones podría revelar insights cualitativos de gran valor, como qué dimensiones ideológicas son más o menos estables en cada modelo, o si las preguntas en las que Qwen censura su respuesta mediante error HTTP 400 son siempre las mismas o dependen del contexto de la sesión. El principal obstáculo para esta línea es, precisamente, el coste de tokens de API descrito en las limitaciones: su viabilidad requiere disponer de presupuesto específicamente asignado o de acceso a recursos computacionales institucionales.

La segunda línea sería ampliar la muestra de modelos evaluados, incluyendo modelos de origen árabe, sudasiático o africano para explorar si el patrón de convergencia progresista se mantiene en modelos de contextos culturales aún más alejados del contexto occidental que da forma a los cuestionarios utilizados. La tercera sería desarrollar o adaptar instrumentos de medición ideológica que no estén anclados exclusivamente en las categorías del espectro político occidental, de modo que permitan capturar dimensiones ideológicas relevantes en contextos como el chino o el árabe sin depender de marcos conceptuales ajenos. La cuarta sería extender el análisis al comportamiento del modelo en interacciones prolongadas, evaluando si el efecto de sycophancy documentado por Nehring et al. (2024) amplifica el posicionamiento ideológico inicial en conversaciones sostenidas con usuarios de distintas orientaciones políticas. Finalmente, una línea de especial relevancia social sería evaluar el impacto real del sesgo ideológico de los LLMs en las opiniones de sus usuarios mediante estudios experimentales de diseño similar al de Bai et al. (2025), pero con medición del posicionamiento previo del usuario y seguimiento longitudinal de su evolución tras el uso continuado de los modelos.

En síntesis, el presente trabajo muestra que los grandes modelos de lenguaje de última generación no son herramientas ideológicamente neutras. Tienen posicionamientos medibles, consistentes y sistemáticamente desplazados hacia el lado progresista del espectro político, con diferencias de grado entre modelos pero sin inversiones de orientación. En un contexto en el que estas herramientas son utilizadas por cientos de millones de personas para informarse, tomar decisiones y formarse opiniones, la comprensión y la transparencia sobre sus sesgos ideológicos constituye una cuestión de relevancia tanto académica como democrática.

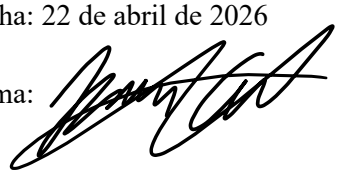
Declaración de Uso de Herramientas de Inteligencia Artificial Generativa

Por la presente, yo, **Javier Gil de Peñaranda**, estudiante de **ADE+BA** de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "**Posicionamiento ideológico en modelos de inteligencia artificial generativa: un análisis comparativa mediante cuestionarios de ideología política**", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Referencias**: Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
2. **Interpretador de código**: Para realizar análisis de datos preliminares.
3. **Corrector de estilo literario y de lenguaje**: Para mejorar la calidad lingüística y estilística del texto.
4. **Generador previo de diagramas de flujo y contenido**: Para esbozar diagramas iniciales.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 22 de abril de 2026

Firma: 

Bibliography

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 337-351.

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). Which humans? Analyzing AI's skewed sense of humanity. *arXiv preprint*.

Bai, H., Voelkel, J. G., Muldowney, S., Eichstaedt, J. C., & Willer, R. (2025). LLM-generated messages can persuade humans on policy issues. *Nature Communications, 16*, 6037.

Bang, Y., Chen, N., Lee, J., & Fung, P. (2024). Do LLMs exhibit political bias? *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics.* Bangkok.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 610–623.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., . . . Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint*.

Buyl, M. e. (2025). Large language models reflect the ideology of their creators. *npj Artificial Intelligence*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *356*(6334), 183-186.

Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023). From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics.* Toronto.

Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv, 2301.01768*.

Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: Measuring ChatGPT political bias. *Public Choice, 198*, 3-23.

Nehring, A. e. (2023). Revisiting the political biases of ChatGPT. *Frontiers in Artificial Intelligence, 6*.

Nehring, A. e. (2024). LLMs as chatbots: Echo chambers and sycophancy. *Proceedings of LREC-COLING 2024.* Torino: ELRA / ACL.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., . . . Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS 2022)*, *35*, pp. 27730–27744.

Pacheco, A. G., Cavalini, A., & Comarela, G. (2025). Echoes of power: Investigating geopolitical bias in US and China large language models. *Humanities and Social Sciences Communications*.

Röttger, P., Vidgen, B., Hovy, D., & Pierrehumbert, J. (2024). Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in large language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics.* Bangkok.

Rozado, D. (2024). The political preferences of LLMs. *PLOS ONE, 19*(7).

Santurkar, S., Durmus, E., Ladd, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *Proceedings of the 40th International Conference on Machine Learning (ICML).* Honolulu.

Simmons, G. (2023). Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *PLOS ONE*.

Zhu, Q., Lyu, D., Fan, X., Wang, X., Tu, Q., Zhan, Y., & Chen, H. (2024). Multi-Model Consistency for LLMs' Evaluation. *IEEE*.