ENCUENTROS | Revista de Ciencias Humanas, Teoría Social y Pensamiento Crítico pp: 356-367

EL ESTADISTA ALGORÍTMICO: ¿PUEDEN LOS MODELOS GRANDES DE LENGUAJE TRANSFORMAR EL ANÁLISIS DE POLÍTICAS PÚBLICAS?

The Algorithmic Statistician: Can Large Language Models Transform Policy Analysis?

Daniel José Boza Muñoz
Universidad Católica Andrés Bello, Venezuela.
https://orcid.org/0009-0008-4528-156X

Este trabajo está depositado en Zenodo:
DOI: https://doi.org/10.5281/zenodo.17307063

RESUMEN

Los Modelos Grandes de Lenguaje (LLMs) se perfilan como herramientas transformadoras para la gobernanza, pero su valor empírico en el análisis de políticas públicas está poco examinado. Mediante una revisión sistemática de 141 publicaciones (Google Scholar, arXiv), este estudio analiza la capacidad de los LLMs para simular agentes heterogéneos, extraer texto jurídico-político y apoyar la toma de decisiones. Al sintetizar avances en modelado multi-agente, extracción neuro-simbólica y plataformas de IA cooperativa, el artículo demuestra que: a) la heterogeneidad de los LLMs aproxima la diversidad demográfica más fielmente que los modelos representativos; b) los sistemas híbridos logran una precisión de última generación en el análisis de documentos; y c) los marcos abiertos, aunque fomentan la transparencia y participación, conllevan riesgos de sesgo y opacidad. Se propone una agenda para mejorar la inferencia multimodal y la gobernanza ética. En conclusión, los LLMs son complementos poderosos—no autosuficientes—del juicio experto, capaces de mejorar el diseño y evaluación de políticas si se integran en ecosistemas sociotécnicos responsables.

Palabras claves: Modelos grandes de lenguaje; análisis de políticas públicas; modelado basado en agentes.

ABSTRACT

Large Language Models (LLMs) are emerging as transformative tools for governance, but their empirical value in public policy analysis remains under-examined. Through a systematic review of 141 publications (Google Scholar, arXiv), this study analyzes the capacity of LLMs to simulate heterogeneous agents, extract legal and political text, and support decision-making. By synthesizing advances in multi-agent modeling, neuro-symbolic extraction, and cooperative AI platforms, the article demonstrates that: a) the heterogeneity of LLMs approximates demographic diversity more accurately than representative models; b) hybrid systems achieve next-generation accuracy in document analysis; and c) open frameworks, while fostering transparency and participation, carry risks of bias and opacity. An agenda for improving multimodal inference and ethical governance is proposed. In conclusion, LLMs are powerful—but not self-sufficient—complements to expert judgment, capable of improving policy design and evaluation if integrated into responsible socio-technical ecosystems.

Keywords: Large language models; public policy analysis; agent-based modeling.

RECIBIDO: 13/02/2025 ACEPTADO: 22/05/2025

INTRODUCCIÓN

En una era de rápido avance tecnológico, los Modelos de Lenguaje de Gran Escala (LLMs, por sus siglas en inglés) son presentados cada vez más como herramientas transformadoras capaces de revolucionar la gobernanza. Desde la simulación de respuestas sociales a políticas fiscales hasta la automatización del análisis de textos legales complejos, el potencial de un "Estadista Algorítmico" para reformar la administración pública parece más cercano que nunca. Sin embargo, en medio del entusiasmo creciente, su valor empírico y utilidad práctica en el dominio matizado del análisis de políticas públicas siguen estando críticamente poco examinados. La pregunta central que anima este estudio es si los LLMs, ya sea de forma individual o como parte de marcos multiagente sofisticados, pueden realmente mejorar o incluso transformar los mecanismos de diseño y evaluación de políticas.

Este artículo aborda esta pregunta mediante una revisión sistemática de la creciente literatura académica, sintetizando hallazgos de 141 publicaciones obtenidas de Google Scholar y arXiv. Al centrarse en avances recientes, este estudio interroga la capacidad de los LLMs para simular agentes heterogéneos, extraer y estructurar datos relevantes para políticas, y generar soporte directo para la toma de decisiones. El análisis demuestra que, aunque los LLMs muestran un potencial notable —ofreciendo aproximaciones más fieles de la diversidad demográfica que los modelos tradicionales y logrando alta precisión en el análisis de documentos—, no son soluciones autosuficientes. Los hallazgos revelan riesgos persistentes de sesgo, opacidad y rezago regulatorio que exigen una gobernanza robusta. En última instancia, este artículo argumenta que los LLMs están mejor posicionados como potentes complementos, en lugar de reemplazos, del juicio experto y la deliberación democrática, desbloqueando su verdadero potencial solo cuando se despliegan dentro de ecosistemas sociotécnicos responsables, transparentes y éticamente fundamentados.

Parámetros de búsqueda y resultados de la revisión sistemática

Esta sección detalla los parámetros de búsqueda sistemática utilizados en dos bases de datos académicas principales—Google Scholar y arXiv—y presenta el conteo inicial de artículos recuperados. El objetivo principal fue identificar trabajos académicos que investigaran la aplicación de los Modelos de Lenguaje de Gran Escala (LLMs) al análisis y a la formulación de políticas públicas. Todas las búsquedas se realizaron asegurando transparencia y reproducibilidad, con criterios de inclusión y exclusión definidos aplicados durante la fase de selección posterior. A continuación, se presenta un resumen de cada búsqueda y del número de documentos inicialmente recuperados.

Búsquedas en Google Scholar

Se realizaron dos búsquedas distintas en Google Scholar para capturar artículos en la intersección entre los LLMs y las políticas públicas.

Primera búsqueda

- Expresión de búsqueda: "large language model" AND "public policy analysis"
- Resultado: 47 artículos recuperados

Segunda búsqueda

- Expresión de búsqueda: "large language model" AND "public policy making"
- Resultado: 48 artículos recuperados

Búsquedas en arXiv

Se realizaron dos búsquedas dirigidas dentro del repositorio arXiv, centradas en artículos del área de Ciencias de la Computación. Estas búsquedas utilizaron términos específicos en el resumen y parámetros de filtrado.

Primera búsqueda

- Parámetros de búsqueda:
- Orden: -announced_date_first
- Tamaño: 50
- Clasificación: Ciencias de la Computación (cs)
- Incluir Cross-List: Sí
- Términos de búsqueda: abstract="large language model" AND abstract="public policy analysis"
- Resultado: 21 artículos recuperados

Segunda búsqueda

- Parámetros de búsqueda:
- Orden: -announced_date_first
- Tamaño: 50
- Clasificación: Ciencias de la Computación (cs)
- Incluir Cross-List: Sí
- Términos de búsqueda: abstract="large language model" AND abstract="public policy making"
- Resultado: 25 artículos recuperados

AGRUPACIÓN INICIAL Y PROCESO DE SELECCIÓN

Tras completar todas las búsquedas en Google Scholar y arXiv, se identificaron inicialmente un total de 141 publicaciones únicas (95 de Google Scholar y 46 de arXiv). Cada uno de estos artículos fue sometido luego a un proceso de selección basado en los siguientes criterios predefinidos:

- Exclusión de estudios centrados principalmente en el análisis de opinión pública (en lugar de análisis o formulación de políticas).
- Exclusión de artículos que no utilizaran o discutieran directamente los Modelos Grandes de Lenguaje.
- Exclusión de artículos que emplearan versiones de Modelos Grandes de Lenguaje lanzad0s antes de 2022.

Los artículos restantes tras este proceso de selección formaron el corpus para la revisión y análisis más detallado en este estudio.

PREGUNTAS Y OBJETIVOS

Este artículo investiga el potencial de los Modelos Grandes de Lenguaje (LLMs) para mejorar el análisis de políticas públicas al servir como proxies escalables de agentes humanos diversos. La pregunta central de investigación es si los LLMs, utilizados de manera individual o dentro de marcos coordinados, pueden superar las limitaciones de los métodos tradicionales de análisis de políticas públicas, incluidos los modelos de agente representativo y la curación manual de datos.

La investigación se basa en trabajos académicos recientes que exploran las capacidades y desafíos del uso de LLMs en contextos relacionados con políticas públicas. Esto incluye aprovechar las diferencias inherentes entre modelos LLM para simular heterogeneidad poblacional (Hao & Xie, 2025), desarrollar marcos de simulación sofisticados para modelar la respuesta pública a intervenciones políticas como campañas de vacunación (Pan et al., 2025), y automatizar la extracción de datos complejos de políticas a partir de textos legales (Buster et al., 2024). El estudio también considera investigaciones que comparan los consejos en políticas generados por LLMs con

los de expertos humanos, revelando diferencias sistemáticas en razonamiento y transparencia (Salekpay et al., 2024), y evalúa el comportamiento de los LLMs en juegos económicos estratégicos para valorar su racionalidad y equidad (Shapira et al., 2024). Además, la investigación incorpora análisis más amplios sobre el impacto de la inteligencia artificial en la política social gubernamental, centrándose en los desafíos institucionales, éticos y sociales de integrar estas tecnologías en la administración pública (Alkhaza'leh, 2025; Lewington et al., 2024; Vatamanu & Tofan, 2025).

En consecuencia, este artículo tiene varios objetivos principales. Busca:

- Evaluar empíricamente la capacidad de marcos de agentes individuales y multi-LLM para replicar razonamientos económicos diversos y respuestas a intervenciones políticas, comparando su desempeño con modelos tradicionales.
- Evaluar métodos para mejorar la validez y transparencia de simulaciones sociales basadas en LLMs, apoyándose en innovaciones recientes como el perfilado demográfico y la modulación de actitudes (Pan et al., 2025; Lewington et al., 2024).
- Comparar sistemáticamente el análisis de políticas generado por LLMs con el consenso experto para identificar y comprender divergencias en sus juicios evaluativos (Salekpay et al., 2024).
- Investigar el potencial y las limitaciones de los LLMs en la automatización de la extracción de datos de políticas, centrándose en la precisión, confiabilidad y escalabilidad (Buster et al., 2024).
- Sintetizar los hallazgos para abordar los desafíos institucionales, técnicos y de gobernanza de integrar LLMs en políticas públicas, proponiendo estrategias para alinear estas herramientas con valores públicos como la transparencia, la rendición de cuentas y la justicia (Alkhaza'leh, 2025; Vatamanu & Tofan, 2025).

Dominio de la política pública y área de aplicación

Un cuerpo emergente de investigación destaca la aplicación creciente de los LLMs en diversos dominios de la política pública, ofreciendo metodologías novedosas para el análisis, al tiempo que presenta desafíos significativos. Una de las principales áreas de aplicación es el uso de LLMs como agentes heterogéneos dentro de modelos basados en agentes (ABMs), para simular respuestas sociales complejas ante intervenciones políticas. Por ejemplo, Hao y Xie (2025) introdujeron un marco multiagente basado en LLMs para analizar la política fiscal, modelando específicamente los impactos conductuales y distributivos de la tributación sobre los ingresos por intereses, al instanciar diferentes LLMs como agentes que representan diversos niveles educativos y tramos de ingresos. De manera similar, en el ámbito de la salud pública, Pan et al. (2025) utilizaron un ABM impulsado por LLMs para simular las respuestas públicas ante intervenciones contra la vacilación vacunal, aprovechando la capacidad de los modelos para generar razonamientos matizados y fundamentados demográficamente. Complementando esto, Shapira et al. (2024) desarrollaron un benchmark para evaluar a los LLMs como agentes autónomos en entornos económicos mediados por lenguaje, como la negociación y el regateo, evaluando su desempeño con métricas relevantes para políticas como la equidad y la eficiencia.

Más allá de la simulación, los LLMs están siendo operacionalizados como herramientas potentes para la extracción, síntesis y análisis de datos con el

fin de apoyar la formulación de políticas basadas en evidencia. En política ambiental, Salekpay et al. (2024) evaluaron críticamente la capacidad de ChatGPT-4 para replicar el consenso experto sobre política climática, concluyendo que los LLMs pueden servir como sintetizadores eficaces, pero no pueden reemplazar el juicio humano matizado. En el ámbito regulatorio de la política energética, Buster et al. (2024) integraron exitosamente LLMs dentro de un marco neuro-simbólico para extraer y estructurar con precisión complejas ordenanzas de zonificación local para proyectos de energía renovable, superando así un importante cuello de botella en la obtención de datos.

Los investigadores también están desarrollando plataformas integradas que aprovechan los LLMs para transformar la creación de políticas y la gobernanza. Lewington et al. (2024) propusieron una plataforma de código abierto que emplea LLMs para la previsión macroeconómica, el análisis del bienestar social y la gobernanza de la IA, incluyendo componentes para la elicitación de valores y una mayor participación ciudadana. Ampliando este enfoque, Alkhaza'leh (2025) analizó el potencial de los LLMs para revolucionar la política social en áreas como la pobreza y la educación, mediante modelos predictivos y optimización de recursos, al tiempo que subrayó los riesgos asociados.

Finalmente, la integración de los LLMs se sitúa dentro del proceso más amplio de transformación digital de la administración pública. Vatamanu y Tofan (2025) evaluaron la adopción de la inteligencia artificial y los LLMs en funciones gubernamentales a lo largo de la Unión Europea.

Hallazgos clave y contribuciones

Los hallazgos se sitúan dentro de, y se ven sustancialmente enriquecidos por, avances recientes en marcos multiagente basados en LLMs (Hao & Xie, 2025; Pan, Nagar y Parkes, 2025), evaluaciones empíricas en dominios sustantivos de política pública (Salekpay, van den Bergh y Savin, 2024), el surgimiento de plataformas cooperativas de formulación de políticas con IA de código abierto (Lewington et al., 2024), la integración de enfoques neuro-simbólicos con LLMs para el análisis de documentos normativos (Buster et al., 2024), análisis multidimensionales del impacto de la IA en la elaboración de políticas sociales (Alkhaza'leh, 2025) y la introducción de benchmarks estandarizados para entornos económicos mediados por lenguaje (Shapira et al., 2024). De manera crucial, estas contribuciones se contextualizan aún más con la rigurosa investigación empírica de Vatamanu y Tofan (2025) sobre la integración de la IA—incluyendo los LLMs—en la administración pública, que pone en primer plano tanto los beneficios medibles como las vulnerabilidades persistentes inherentes a la transformación digital de la gobernanza.

Las distintas investigaciones demuestran que los LLMs exhiben capacidades no triviales de razonamiento económico y conductual, tanto en tareas formales de optimización como en escenarios de políticas más intuitivos y dependientes del contexto. La evaluación empírica entre múltiples LLMs de última generación revela una variación significativa en precisión, consistencia y naturaleza cualitativa de la toma de decisiones. Estas diferencias no son meramente estocásticas, sino que reflejan perfiles cognitivos y conductuales distintos propios de cada modelo (Hao & Xie, 2025; Pan et al., 2025). Por ejemplo, modelos como DeepSeek-V3 y Qwen-2.5 exhiben alta precisión en optimización y razonamiento matizado, mientras que otros, como Llama-3.1-405B y GPT-4o, manifiestan patrones de decisión menos óptimos o más dispersos. Esta heterogeneidad subraya la inadecuación de tratar a los LLMs como cajas negras intercambiables en simulaciones de políticas y destaca la impor-

tancia de la selección y evaluación sistemática de modelos.

La integración de marcos Multi-LLM-Agent-Based (MLAB) representa un avance metodológico frente a los modelos tradicionales basados en agentes y a los enfoques previos con un único LLM. Partiendo de Hao y Xie (2025), este estudio confirma que la heterogeneidad en el comportamiento simulado surge no solo de diferencias en circunstancias objetivas (p. ej., ingreso, educación), sino también de rasgos cognitivos y culturales subjetivos codificados en distintos LLMs. Pan et al. (2025) amplían esta visión mediante el marco VACSIM, que orquesta una población de agentes impulsados por LLMs inicializados con atributos demográficos empíricamente fundamentados e integrados en redes sociales realistas. La capacidad de ajustar las actitudes de los agentes y calibrar las simulaciones con trayectorias conductuales del mundo real (como la vacilación ante las vacunas) demuestra que las sociedades de agentes basadas en LLMs pueden aproximarse tanto a tendencias agregadas como a respuestas específicas de grupo ante intervenciones políticas, siempre que se implementen salvaguardas metodológicas—como la modulación de actitudes y el calentamiento de simulaciones.

La base de evidencia para estas afirmaciones es sólida y multidimensional. Experimentos cuantitativos que emplean tareas económicas y conductuales estandarizadas, múltiples métricas de rendimiento y calibración rigurosa con datos demográficos y de encuestas sustentan la heterogeneidad observada en el comportamiento de los LLMs (Hao & Xie, 2025; Pan et al., 2025). Los análisis cualitativos de los razonamientos generados por los modelos y las narrativas de los agentes iluminan aún más la diversidad de motivos, estilos de razonamiento y sesgos conductuales, incluidas referencias culturales y dinámicas de confianza. Es importante destacar que ambos estudios demuestran que las simulaciones de políticas en marcos multiagente con LLMs generan respuestas específicas por grupo y, en ocasiones, no monótonas a las intervenciones, desafiando la suficiencia de los paradigmas de agente representativo o de modelo único para capturar la diversidad del mundo real. La validez externa de estas simulaciones se ve reforzada por el acuerdo experto con los rankings de políticas (Pan et al., 2025), aunque persisten limitaciones notables, como desalineación demográfica o estocasticidad excesiva en algunos LLMs.

Salekpay et al. (2024) ofrecen una comparación sistemática entre ChatGPT-4 y una encuesta global de expertos, revelando que, si bien los consejos de política generados por LLMs son generalmente informativos, de alta calidad y frecuentemente alineados con el consenso experto, también presentan limitaciones distintivas. Específicamente, ChatGPT demuestra menos sesgo disciplinario que los expertos humanos, lo que contribuye a resultados más holísticos pero a veces insuficientemente discriminatorios. La tendencia del modelo a ponderar por igual todos los criterios de evaluación—en contraste con la priorización de la efectividad y la equidad por parte de los expertos—resalta un riesgo de "corrección política" y falta de selectividad en la síntesis algorítmica. Además, si bien los LLMs sobresalen en la agregación rápida y síntesis de literaturas diversas, su naturaleza de "caja negra" y la falta de transparencia en la procedencia de la información limitan su confiabilidad epistémica e interpretabilidad. Estos hallazgos subrayan que los LLMs pueden servir como valiosos complementos, pero no sustitutos, de procesos dirigidos por expertos en dominios de políticas complejos y cargados de valores. La innovación metodológica de comparar directamente los LLMs con instrumentos de encuestas expertas avanza aún más

la comprensión del campo sobre tanto las promesas como los peligros de la producción algorítmica de conocimiento en contextos de políticas públicas.

El surgimiento de plataformas cooperativas de formulación de políticas públicas con IA de código abierto marca un desarrollo crucial en el campo (Lewington et al., 2024). Lewington et al. (2024) presentan un marco técnicamente sofisticado que integra pronósticos económicos multimodales, generación de políticas alineadas con valores y evaluación empírica de impacto social dentro de una infraestructura transparente, sin fines de lucro y orientada por la comunidad. Su modelo "Economics Transformer" demuestra la viabilidad de fusionar series temporales macroeconómicas y microeconómicas con documentos textuales de políticas, permitiendo un análisis más rico y contextualizado que los enfoques econométricos tradicionales o los LLMs unimodales. La introducción de un componente "AI Legislator" operacionaliza la elicitación y agregación de valores a gran escala mediante modelado bayesiano jerárquico y la Teoría de Fundamentos Morales, fundamentando así las propuestas de política en preferencias públicas empíricamente derivadas y reforzando la legitimidad democrática de las recomendaciones generadas por IA. El compromiso con el desarrollo de código abierto, el benchmarking en vivo (DBITS) y la evaluación empírica establece nuevos estándares de transparencia, rendición de cuentas y mejora continua en el análisis de políticas asistido por IA. Además, la inclusión explícita de evaluación de impacto social—en particular en lo que respecta al mercado laboral y los efectos sobre el bienestar derivados de la adopción de LLMs—aborda una brecha crítica en investigaciones previas, garantizando que las recomendaciones de política respondan a riesgos y oportunidades reales, y no solo a consideraciones técnicas o teóricas. Sin embargo, el estado preliminar de la plataforma y los desafíos para escalar la elicitación de valores y asegurar representatividad resaltan obstáculos metodológicos y prácticos aún vigentes.

La integración de enfoques neuro-simbólicos con LLMs para el análisis de documentos normativos, como lo ejemplifica Buster et al. (2024), representa un avance metodológico y práctico significativo. Buster et al. (2024) demuestran que la combinación de LLMs con marcos de árboles de decisión—incorporando así conocimiento experto temático y lógica simbólica—permite una extracción altamente precisa (85–90%) y escalable de datos estructurados de políticas a partir de textos legales heterogéneos, como las ordenanzas de ubicación de energía eólica. Este enfoque híbrido supera tanto a los métodos de un solo prompt como a las estrategias de generación aumentada por recuperación (RAG), que enfrentan dificultades ante la complejidad y variabilidad de los documentos legales. La automatización resultante no solo reduce la carga laboral y el error humano, sino que también respalda el mantenimiento de bases de datos de políticas actualizadas a escala nacional, cruciales para la investigación en políticas energéticas y la modelación de escenarios. Es importante que Buster et al. (2024) realicen una evaluación crítica de las limitaciones de los LLMs en contextos legales y normativos, enfatizando la necesidad de supervisión humana, validación rigurosa y salvaguardas contextuales específicas para mitigar riesgos como alucinaciones o clasificaciones erróneas. Su publicación en código abierto y la documentación metodológica detallada refuerzan aún más la reproducibilidad y adopción, mientras que la demostrada generalización del enfoque señala su aplicabilidad potencial en diversos dominios de política. No obstante, el estudio subraya que los LLMs deben complementar, y no

reemplazar, el juicio experto y la revisión legal, y que el análisis de errores y la transparencia rigurosa siguen siendo esenciales.

La introducción de GLEE—un marco y benchmark unificado para entornos económicos mediados por lenguaje—por Shapira et al. (2024) constituye un avance fundamental en la evaluación empírica de LLMs para tareas relevantes en política pública. GLEE modela y simula sistemáticamente juegos económicos secuenciales de dos jugadores mediados por lenguaje, incluyendo negociación, regateo y persuasión, en una amplia gama de parámetros de mercado y modalidades de comunicación. Shapira et al. (2024) demuestran que la eficiencia, equidad y beneficio propio en estas interacciones no solo son altamente sensibles a la estructura de la información y a la comunicación lingüística, sino también contingentes al emparejamiento específico de agentes y configuraciones del juego. Es notable que ningún LLM supere consistentemente a los demás en todos los escenarios, y que los participantes humanos frecuentemente exhiban comportamientos más extremos o sesgados—como el anclaje—que los LLMs. La presencia y forma del lenguaje natural alteran fundamentalmente las dinámicas estratégicas, a menudo de maneras que divergen de la teoría económica clásica. El benchmarking estandarizado y de código abierto de GLEE permite una investigación robusta, reproducible y comparativa, cerrando la brecha entre la abstracción económica/teórica del juego y la complejidad lingüística de la deliberación política real. La capacidad del marco para análisis comparativos a gran escala entre humanos e IA proporciona un fundamento empírico crítico para el diseño de sistemas de decisión híbridos y la evaluación de la racionalidad, equidad y comportamiento social de los LLMs en contextos de políticas públicas. Al revelar la ausencia de un LLM óptimo universal y la naturaleza contextual de las interacciones entre agentes, Shapira et al. (2024) cuestionan las suposiciones de universalidad de modelo y destacan la necesidad de un despliegue y evaluación matizados y conscientes del contexto de los LLMs en el análisis de políticas públicas.

El presente análisis se ve sustancialmente informado y ampliado por la investigación exhaustiva de Alkhaza'leh (2025) sobre el impacto de la IA en la formulación de políticas sociales. Alkhaza'leh (2025) demuestra que los LLMs y las tecnologías de IA relacionadas pueden mejorar la efectividad y eficiencia de la política social al permitir un análisis sofisticado de conjuntos de datos complejos y a gran escala, mejorando así la identificación de necesidades sociales, la priorización de intervenciones y la asignación de recursos. El estudio también destaca que la integración de IA puede promover la transparencia y objetividad en los procesos de política, reduciendo potencialmente el sesgo humano y fomentando la confianza pública en las instituciones gubernamentales. Cabe destacar el énfasis en la participación cívica: se demuestra que las plataformas impulsadas por IA, incluidos chatbots y foros digitales basados en LLMs, facilitan una participación ciudadana más amplia y democratizan la deliberación sobre políticas. Alkhaza'leh (2025) también demuestra la utilidad de la IA en la supervisión dinámica y la respuesta a crisis, con aplicaciones en análisis de tendencias en tiempo real y modelación predictiva para una acción gubernamental rápida durante emergencias. Sin embargo, la investigación evalúa críticamente la naturaleza dual de la IA, reconociendo riesgos como el desplazamiento laboral, el agravamiento de desigualdades sociales y la persistencia de desafíos éticos y regulatorios—especialmente en relación con la privacidad, el sesgo y la equidad. El marco estructurado del estudio para la integración de

la IA en todas las etapas del ciclo de políticas, el análisis sectorial específico (por ejemplo, pobreza, desempleo, educación, salud), y las recomendaciones prácticas para supervisión ética e inversión en alfabetización digital proporcionan tanto un andamiaje teórico como práctico para una adopción responsable de la IA. Es importante que la síntesis de Alkhaza'leh (2025) entre evidencia empírica, argumentación analítica y reflexión normativa destaque la necesidad de marcos regulatorios robustos y gobernanza interdisciplinaria para asegurar que la innovación en políticas impulsada por IA se alinee con principios de justicia, equidad y responsabilidad.

Vatamanu y Tofan (2025) proporcionan evidencia empírica rigurosa de que la digitalización impulsada por IA—operacionalizada mediante el Índice de Economía y Sociedad Digital (DESI)—está positivamente asociada con la calidad de la gobernanza y el crecimiento económico en los Estados miembros de la UE. Su análisis demuestra que la integración de servicios públicos digitales y tecnologías avanzadas, incluida la analítica habilitada por IA, es un predictor estadísticamente significativo de mejoras en eficiencia administrativa, transparencia y participación ciudadana. Sin embargo, Vatamanu y Tofan (2025) también identifican vulnerabilidades persistentes y multifacéticas que acompañan la adopción de IA en la administración pública, incluidos el sesgo algorítmico, los riesgos de ciberseguridad, los desafíos de adaptación laboral y los dilemas éticos. Estas vulnerabilidades, si no se abordan adecuadamente, amenazan con socavar la confianza pública, la equidad y la legitimidad de la gobernanza digital. El marco analítico multidimensional de los autores—que sintetiza teorías de adopción tecnológica, modelos de gobernanza de IA e indicadores empíricos—ofrece una lente integral para comprender tanto las oportunidades como los riesgos de la integración de LLMs. Sus hallazgos refuerzan la necesidad de marcos legales y éticos sólidos, participación pública continua e inversiones dirigidas en infraestructura digital y habilidades para garantizar que los beneficios de la transformación impulsada por IA se materialicen de forma inclusiva y sostenible. De manera notable, Vatamanu y Tofan (2025) moderan el tecno-optimismo predominante al poner en primer plano las complejidades organizativas, legales y éticas de la implementación, y al ofrecer recomendaciones de política directamente relevantes para la expansión responsable de los LLMs en contextos de políticas públicas.

Desafíos, limitaciones y riesgos identificados

Los LLMs siguen siendo fundamentalmente "cajas negras" opacas, lo que dificulta la trazabilidad, la auditoría y la rendición de cuentas pública (Buster et al., 2024; Salekpay et al., 2024; Vatamanu & Tofan, 2025). El sesgo algorítmico—arraigado en desigualdades históricas presentes en los datos de entrenamiento—amenaza con producir resultados discriminatorios en políticas, especialmente en dominios sensibles como el bienestar social y la justicia penal (Alkhaza'leh, 2025; Pan et al., 2025). La heterogeneidad de los textos legales, combinada con los estrictos límites de la ventana de contexto, dificulta la extracción precisa y el razonamiento, requiriendo con frecuencia preprocesamientos propensos a errores (Buster et al., 2024). Las preocupaciones en torno a la privacidad de los datos, la ciberseguridad y el cumplimiento normativo son prominentes (Lewington et al., 2024; Vatamanu & Tofan, 2025), mientras que la adaptación de la fuerza laboral, la resistencia organizacional y la confianza social condicionan las trayectorias de adopción (Alkhaza'leh, 2025). Otros desafíos incluyen la ambigüedad regulatoria (Vatamanu & Tofan, 2025), las alucinaciones y omisiones como modos de

fallo (Buster et al., 2024), el alto consumo energético (Lewington et al., 2024), los ecosistemas de datos fragmentados (Lewington et al., 2024), el comportamiento inconsistente de los agentes (Hao & Xie, 2025; Pan et al., 2025), la escasa cuantificación de la incertidumbre (Lewington et al., 2024), la sensibilidad a los prompts (Salekpay et al., 2024) y la ausencia de benchmarks estandarizados de evaluación (Shapira et al., 2024).

El alcance empírico es limitado: los estudios suelen examinar poblaciones reducidas de agentes, dominios de política únicos o conjuntos de datos regionales, lo que restringe su generalización (Hao & Xie, 2025; Lewington et al., 2024). Los resultados de los LLMs dependen en gran medida de la síntesis cualitativa en lugar de evidencia cuantitativa actual, lo que reduce el rigor empírico (Salekpay et al., 2024). La opacidad en las fuentes introduce sesgos disciplinares incontrolables, y los indicadores por aproximación corren el riesgo de incurrir en endogeneidad (Vatamanu & Tofan, 2025). El rendimiento del modelo varía drásticamente según el diseño del prompt, los hiperparámetros y la arquitectura elegida, lo que socava su robustez (Pan et al., 2025). Además, la colaboración en código abierto—aunque mejora la transparencia—no resuelve por completo las restricciones éticas, legales o específicas del contexto (Lewington et al., 2024).

Finalmente, una dependencia no controlada de las recomendaciones generadas por LLMs puede distorsionar el discurso público, fomentar el sesgo de automatización y erosionar la gobernanza deliberativa (Salekpay et al., 2024). La amplificación del sesgo puede marginar a poblaciones vulnerables (Pan et al., 2025), mientras que la limitada explicabilidad desafía la supervisión democrática (Buster et al., 2024). Las vulnerabilidades persistentes en la privacidad de los datos y el aumento de superficies de ataque amenazan la seguridad nacional y la autonomía individual (Lewington et al., 2024; Vatamanu & Tofan, 2025). Los riesgos a nivel social incluyen el rezago regulatorio, el desplazamiento laboral y los costos medioambientales (Alkhaza'leh, 2025). De forma más fundamental, delegar juicios de política cargados de valores a modelos opacos plantea interrogantes sobre la agencia humana y el pluralismo (Shapira et al., 2024).

Direcciones futuras de investigación propuestas

La literatura actual reconoce el potencial de los modelos grandes de lenguaje (LLMs) para el análisis de políticas públicas, pero destaca brechas persistentes que requieren una investigación sistemática. Ante todo, los modelos basados en agentes (ABMs) suelen depender de variaciones a nivel de prompt de un único LLM, lo que subrepresenta los rasgos cognitivos, culturales y socioeconómicos heterogéneos que existen en las poblaciones reales (Hao & Xie, 2025; Pan, Nagar y Parkes, 2025). Los trabajos futuros deben desarrollar técnicas de inicialización y alineación que generen agentes más fieles y coherentes demográficamente, al tiempo que amplían los ABMs para capturar influencias dinámicas entre pares, aprendizaje social y fenómenos emergentes como la polarización y la formación de normas (Pan et al., 2025; Shapira et al., 2024).

La robustez metodológica sigue siendo un tema de preocupación: los resultados de las simulaciones varían significativamente entre arquitecturas de LLM y diseños de prompt, y muchos modelos exhiben sesgos que amenazan su capacidad de generalización—especialmente en el caso de grupos subrepresentados (Hao & Xie, 2025; Pan et al., 2025). Por ello, es esencial contar con benchmarking riguroso entre modelos, validaciones longitudinales con datos humanos y estándares transparentes de reporte. Los riesgos éticos y de representa-

ción, incluidos razonamientos opacos, agregación indiscriminada de fuentes y respuestas inestables, refuerzan aún más la necesidad de supervisión interdisciplinaria (Salekpay, van den Bergh y Savin, 2024).

La integración de evidencia multimodal constituye otra frontera. Las series temporales económicas y los documentos textuales de políticas siguen estando compartimentalizados, lo que limita la inferencia causal y la precisión en las proyecciones. Nuevas arquitecturas—como transformadores con valores continuos y atención cruzada entre modalidades—deberían permitir un razonamiento conjunto sobre texto y datos numéricos, respaldado por rankings abiertos y dinámicos que evolucionen con indicadores económicos en tiempo real (Lewington et al., 2024). De forma complementaria, se debe automatizar la recuperación de documentos normativos, mejorar la precisión de extracción en distintos dominios regulatorios e incorporar detección sistemática de errores con revisión humana en el ciclo (Buster et al., 2024).

La agenda de investigación también abarca seguridad, privacidad, transparencia y alineación de valores. Diversos autores abogan por pipelines automatizados pero interpretables que puedan captar valores públicos heterogéneos, integrarlos en la generación de políticas y democratizar el acceso mediante diseños centrados en el usuario (Lewington et al., 2024; Alkhaza'leh, 2025). Estudios comparativos entre sectores y jurisdicciones deben evaluar los impactos sobre la equidad, el mercado laboral y los resultados de gobernanza, prestando especial atención a la mitigación de sesgos y a los marcos de rendición de cuentas (Alkhaza'leh, 2025; Vatamanu & Tofan, 2025).

Plataformas de evaluación estandarizadas como el marco GLEE ofrecen un camino hacia experimentos reproducibles en entornos estratégicos mediados por lenguaje y deben adaptarse a escenarios relevantes para políticas públicas (Shapira et al., 2024).

En conjunto, estas direcciones abogan por: (a) una mayor heterogeneidad de agentes y modelado de interacciones; (b) benchmarking riguroso y transparente entre LLMs; (c) arquitecturas multimodales y conjuntos de datos dinámicos; (d) generación de políticas alineada con valores y fundada éticamente; y (e) investigación en el sector público sensible al contexto. Abordar estas prioridades promete simulaciones de políticas impulsadas por LLMs más realistas, equitativas y aplicables, avanzando tanto en la comprensión académica como en la gobernanza práctica.

REFERENCIAS

Hao, Y., & Xie, D. (2025, February 24). A multi-LLM-agent-based framework for economic and public policy analysis (arXiv:2502.16879). *arXiv*. https://arxiv.org/abs/2502.16879

Pan, S., Nagar, S., & Parkes, D. C. (2025). Can a society of generative agents simulate human behavior and inform public health policy? A case study on vaccine hesitancy (Version 3; arXiv:2503.09639). *arXiv*. https://arxiv.org/abs/2503.09639

Salekpay, F., van den Bergh, J., & Savin, I. (2024). Comparing advice on climate policy between academic experts and ChatGPT. *Ecological Economics, 226*, Article 108352. https://doi.org/10.1016/j.ecolecon.2024.108352

Lewington, A., et al. (2024). Creating a cooperative AI policymaking platform through open source collaboration. *arXiv*. https://doi.org/10.48550/arXiv.2412.06936

Buster, G., Pinchuk, P., Barrons, J., McKeever, R., Levine, A., & Lopez, A. (2024). Supporting energy policy research with large language models. *arXiv*. https://doi.org/10.48550/arXiv.2403.12924

Alkhaza'leh, Y. M. S. (2025). Artificial intelligence and its impact on social policy-making. *Journal of Ecohumanism, 4*(1), 2320–2337. https://doi.org/10.62754/joe.v4i1.6053

Shapira, E., Madmon, O., Reinman, I., Amouyal, S. J., Reichart, R., & Tennenholtz, M. (2024). GLEE: A unified framework and benchmark for language-based economic environments (arXiv:2410.05254). *arXiv*. https://arxiv.org/abs/2410.05254

Vatamanu, A. F., & Tofan, M. (2025). Integrating artificial intelligence into public administration: Challenges and vulnerabilities. *Administrative Sciences, 15*(4), Article 149. https://doi.org/10.3390/admsci15040149