Propuesta de Trabajo Final de Máster

Sistema Multiagente para Verificación Automática de Hechos en Videos de YouTube

Begoña Echavarren Sánchez

Tutor: Josep-Anton Mir Tutusaus

Máster Universitario en Ciencia de Datos
Universitat Oberta de Catalunya

PEC 1 - Definición del TFM

Octubre 2025

Índice

1. Título del proyecto

Sistema Multiagente para Verificación Automática de Hechos en Videos de YouTube

2. Palabras clave

fact-checking automatizado, verificación de hechos, LLMs, modelos de lenguaje, sistemas multiagente, procesamiento de lenguaje natural, desinformación, recuperación de información, análisis de evidencias

3. Resumen

La desinformación en plataformas de video constituye un desafío creciente en la sociedad digital actual. Este trabajo propone el diseño, implementación y evaluación de un sistema end-to-end de verificación automática de hechos (fact-checking) para videos de YouTube, basado en una arquitectura multiagente que integra modelos de lenguaje de gran escala (LLMs), técnicas de procesamiento de lenguaje natural y recuperación de información en línea.

El sistema propuesto opera mediante un pipeline de cinco componentes especializados: (1) extracción de transcripciones de video, (2) identificación automática de afirmaciones factuales mediante LLMs, (3) generación de consultas de búsqueda optimizadas, (4) recuperación y evaluación de evidencias desde fuentes web, y (5) síntesis de veredictos razonados con evaluación de confianza. Cada componente procesa información estructurada garantizando robustez y trazabilidad en todo el proceso.

La implementación técnica se desarrollará en Python utilizando frameworks modernos como Pydantic AI, FastAPI, entre otros, junto con modelos de lenguaje tanto comerciales como de código abierto. El proyecto incluirá una interfaz web desarrollada en React con streaming en tiempo real mediante Server-Sent Events (SSE), proporcionando una experiencia de usuario interactiva durante el proceso de verificación.

Un componente fundamental del trabajo será la evaluación comparativa de diferentes LLMs y enfoques arquitecturales. Se implementarán técnicas de evaluación específicas para modelos de lenguaje, métricas de rendimiento cuantitativas y protocolos de evaluación cualitativa para determinar las configuraciones óptimas del sistema. El análisis incluirá estudios de caso en diferentes temáticas de videos (economía, geopolítica, ciencia) con especial atención a los compromisos entre coste, calidad y latencia en las distintas aproximaciones evaluadas.

4. Descripción y justificación del proyecto

4.1. Contexto del problema

La desinformación en plataformas de video constituye un desafío creciente en la sociedad digital actual. YouTube, con más de 2 mil millones de usuarios activos mensuales, se ha convertido en una fuente primaria de información para millones de personas. Sin embargo, la ausencia de mecanismos efectivos de verificación de información permite la propagación de afirmaciones incorrectas o engañosas a escala masiva.

El problema central que aborda este proyecto es la **falta de mecanismos automatizados y accesibles para verificar afirmaciones factuales en contenido audiovisual de plataformas digitales**, específicamente videos de YouTube. Los usuarios individuales carecen de herramientas para verificar afirmaciones de forma sistemática, mientras que la verificación manual requiere habilidades avanzadas, tiempo considerable y recursos no siempre disponibles.

4.2. Relevancia e importancia

Este proyecto es relevante por múltiples razones:

- **Impacto social:** La desinformación erosiona la confianza en instituciones, influye en decisiones políticas y afecta la salud pública [2, 30]. Un sistema automatizado de verificación puede contribuir a mitigar estos efectos.
- **Desafío tecnológico:** El proyecto integra múltiples áreas de investigación en ciencia de datos: procesamiento de lenguaje natural, recuperación de información, evaluación de credibilidad de fuentes y razonamiento automático. Representa un problema técnicamente complejo que requiere soluciones innovadoras.
- **Escalabilidad necesaria:** El volumen de contenido generado diariamente (más de 500 horas de video por minuto en YouTube) hace inviable el fact-checking puramente humano. La automatización es necesaria para abordar el problema a escala.
- **Alfabetización mediática:** Herramientas que empoderan a ciudadanos para evaluar críticamente información promueven el pensamiento crítico y fortalecen la alfabetización mediática en la sociedad.

4.3. Estado actual de soluciones existentes

Actualmente existen tres enfoques principales para la verificación de hechos:

1. **Fact-checking manual:** Organizaciones como Newtral, Maldita.es y FactCheck.org emplean periodistas especializados. Este enfoque garantiza calidad pero está limitado por recursos humanos y es inherentemente lento.

2. **Sistemas académicos:** Proyectos como FEVER [26], FEVEROUS [1] y trabajos recientes sobre verificación científica [28] operan principalmente sobre bases de conocimiento estructuradas (Wikipedia) o requieren datasets curados específicos. No están disponibles como productos utilizables por usuarios finales y tienen alcance limitado.

3. **APIs de verificación:** Google Fact Check Tools API agrega fact-checks ya publicados por organizaciones profesionales, pero solo cubre contenido previamente verificado manualmente. No proporciona verificación nueva de contenido no examinado.

Limitaciones identificadas: Ninguno de estos enfoques proporciona verificación automatizada end-to-end para contenido audiovisual generado continuamente por usuarios en plataformas como YouTube. Existe una brecha entre la investigación académica (datasets estáticos, dominios limitados) y herramientas prácticas accesibles para usuarios finales.

4.4. Propuesta y contribución esperada

Este proyecto desarrollará un **sistema funcional end-to-end** que procese videos de YouTube mediante:

- Extracción automática de transcripciones
- Identificación de afirmaciones factuales verificables mediante LLMs
- Recuperación de evidencias desde múltiples fuentes web
- Evaluación de credibilidad de fuentes y postura de evidencias
- Generación de veredictos fundamentados con explicaciones transparentes
- Interfaz web accesible para usuarios finales

Contribuciones esperadas:

1. **Arquitectura multiagente modular y escalable:** Diseño e implementación de un sistema multiagente optimizado compuesto por cinco componentes especializados que operan de forma coordinada mediante esquemas de datos estructurados (Figura 1). Esta arquitectura modular permite procesamiento paralelo de múltiples afirmaciones, facilita la optimización y extensibilidad del sistema y garantiza trazabilidad completa del pipeline de verificación.

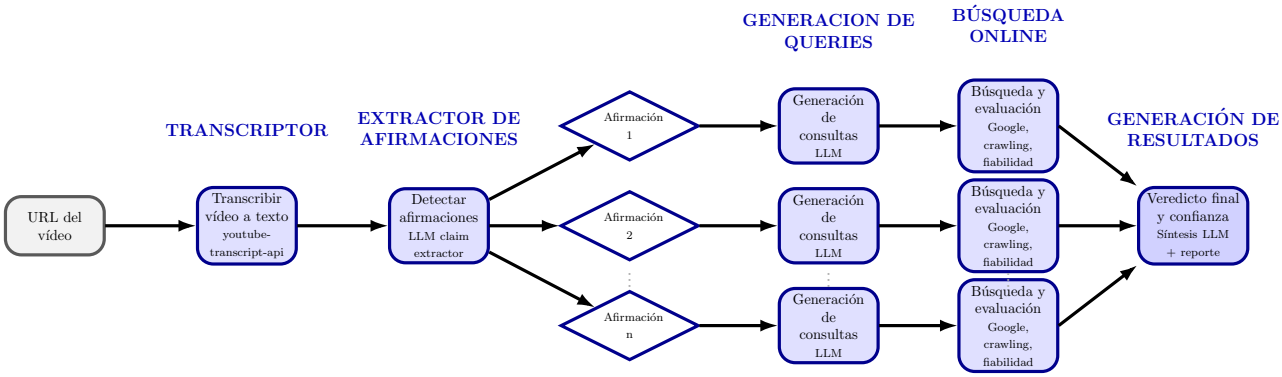

Figura 1: Arquitectura del sistema multiagente propuesto. Pipeline de cinco componentes especializados procesando secuencialmente desde transcripción hasta veredicto final.

2. **Sistema completo operativo:** A diferencia de prototipos académicos, se desarrollará una aplicación funcional con interfaz web.

3. **Análisis comparativo de LLMs:** Evaluación de modelos comerciales (i.e. GPT, Claude) versus open-source (i.e. Qwen, Deepseek), documentando compromisos entre coste, rendimiento y calidad.

4. **Metodología de evaluación:** Framework de evaluación cualitativa y cuantitativa aplicable a sistemas similares.

5. **Consideración ética y de sesgos:** Análisis cualitativo de sesgos en el procesamiento y uso de los modelos que minimicen los riesgos de la automatización del proceso de verificación.

5. Motivación personal

Mi interés en este proyecto surge de la confluencia de tres factores principales:

Experiencia profesional: Como Machine Learning Engineer, trabajo en proyectos de NLP, APIs y sistemas productivos. Este TFM me permite aplicar y profundizar conocimientos en un contexto de alto impacto social, mientras exploro tecnologías emergentes como LLMs y arquitecturas multiagente.

Intereses personales: Soy consumidor habitual de contenido educativo en YouTube sobre economía, geopolítica y ciencia. Con frecuencia encuentro afirmaciones que requieren verificación pero carezco de herramientas eficientes para hacerlo sistemáticamente. Este proyecto nace de una necesidad personal real: crear una herramienta que me gustaría tener como usuario.

Objetivos de desarrollo: El proyecto combina aspectos de investigación (estado del arte en fact-checking, evaluación de LLMs) con ingeniería de software aplicada

(arquitectura escalable, despliegue de aplicaciones). Esta dualidad se alinea con mi visión profesional de combinar investigación aplicada y desarrollo de producto.

Adicionalmente, considero que el fact-checking automatizado representa una contribución al bien social. En una era donde la desinformación tiene consecuencias reales en salud pública, democracia y cohesión social, desarrollar herramientas que fomenten el pensamiento crítico y la alfabetización mediática es una responsabilidad que me motiva profundamente.

6. Objetivos del proyecto

6.1. Hipótesis u objetivo principal

Diseñar, implementar y evaluar un sistema multiagente basado en modelos de lenguaje capaz de verificar automáticamente afirmaciones factuales en videos de YouTube mediante recuperación y análisis de evidencias web, generando veredictos fundamentados con estimación de confianza, que sean útiles para usuarios finales.

6.2. Objetivos específicos (preguntas de investigación)

1. **Diseño de arquitectura multiagente**
   - Definir el pipeline de procesamiento de forma modular con componentes especializados.
   - Establecer esquemas de datos estructurados para la comunicación entre agentes.
   - Diseñar el sistema de orquestación de la información.
2. **Implementación de componentes del sistema**
   - Desarrollar el módulo de extracción de transcripciones de vídeos desde YouTube.
   - Implementar el agente de identificación de claims con clasificación de relevancia.
   - Crear el generador de consultas de búsqueda optimizadas.
   - Desarrollar el sistema de web scraping con evaluación de fiabilidad de fuentes.
   - Implementar el generador de veredictos explicables y trazables.
3. **Optimización del uso de LLMs**
   - Comparar el rendimiento entre modelos comerciales (p. ej. GPT-4o-mini) y open-source (p. ej. Qwen 3, DeepSeek).

- Analizar compromisos entre coste, latencia y calidad de outputs.
- Evaluar estrategias de prompt optimization y structured outputs.

4. **Evaluación del sistema**
   - Definir métricas de rendimiento técnico: tiempo de respuesta, coste por token, throughput y tasa de finalización por agente.
   - Incluir métricas cuantitativas específicas de LLMs como faithfulness, factual consistency, coherence y stance accuracy.
   - Incorporar evaluación cuantitativa y cualitativa con LLM-as-a-judge y revisión manual.
   - Realizar estudios de caso en distintas temáticas y analizar limitaciones del sistema.
5. **Implementación de sistema end-to-end**
   - Desarrollar una API REST con streaming en tiempo real.
   - Crear una interfaz web intuitiva para usuarios finales.
   - Garantizar reproducibilidad y generar documentación.
6. **Consideración de implicaciones éticas y de sesgo**
   - Analizar los sesgos potenciales en las fuentes y en los modelos utilizados.
   - Identificar limitaciones del enfoque automatizado, su posible impacto en la propagación de desinformación, y definir buenas prácticas que mitiguen esos riesgos durante la verificación.

6.3. Alcance del proyecto

Incluye:

- Videos de YouTube con transcripciones disponibles (subtítulos automáticos o manuales)
- Afirmaciones factuales verificables mediante fuentes web públicas
- Contenido en inglés
- Procesamiento bajo demanda de videos individuales

No incluye:

- Generación propia de transcripciones desde audio (speech-to-text)

- Verificación de contenido visual
- Verificación a gran escala (batch)
- Contenido de plataformas distintas a YouTube

7. Consideraciones Éticas y de Impacto Social

7.1. Competencia de Compromiso Ético y Global (CCEG)

El proyecto aborda las tres dimensiones de la CCEG establecidas por la UOC:

(I) Sostenibilidad: El uso de LLMs implica consumo energético significativo. Se abordará mediante evaluación de modelos open-source locales versus comerciales cloud, optimización de prompts para minimizar tokens procesados, y documentación transparente del consumo de recursos.

(II) Comportamiento ético y responsabilidad social: Los LLMs pueden perpetuar sesgos presentes en sus datos de entrenamiento. Estrategias de mitigación: transparencia total en fuentes consultadas con URLs verificables, presentación de evidencias mostrando diferentes posturas (apoya/refuta/neutral), evaluación explícita de nivel de confianza en cada veredicto, y advertencias claras sobre limitaciones del sistema automatizado. El sistema se posiciona como herramienta de apoyo que automatiza pasos preliminares, permitiendo a profesionales enfocarse en análisis complejos que requieren juicio humano.

(III) Diversidad y derechos humanos: El sistema busca evaluar diversidad de perspectivas cuando múltiples fuentes web con diferentes enfoques estén disponibles, documentando esta limitación cuando no sea posible. No censura contenido sino que proporciona contexto adicional mediante verificación, respetando la libertad de expresión. No procesa datos personales ni información identificable de usuarios.

7.2. Objetivos de Desarrollo Sostenible (ODS)

El proyecto contribuye a tres ODS de la Agenda 2030:

ODS 4 (Educación de Calidad): El sistema actúa como herramienta de alfabetización mediática que promueve pensamiento crítico y capacidad de evaluación de fuentes de información. Facilita el aprendizaje sobre verificación de hechos, análisis de evidencias y reconocimiento de afirmaciones verificables, competencias fundamentales en la sociedad digital actual donde el volumen de información supera la capacidad individual de validación.

ODS 16 (Paz, Justicia e Instituciones Sólidas): Combate la desinformación que erosiona la confianza en instituciones democráticas y medios de comunicación. Proporciona mecanismos de verificación transparentes, trazables y accesibles que permiten

a ciudadanos contrastar afirmaciones factuales de forma sistemática, contribuyendo a un ecosistema informativo más saludable y resiliente ante campañas de desinformación.

ODS 10 (Reducción de Desigualdades): Democratiza el acceso a verificación de hechos mediante herramienta gratuita y de código abierto, reduciendo la brecha entre organizaciones profesionales con recursos especializados (fact-checkers, periodistas) y ciudadanos individuales. Empodera a usuarios sin formación especializada para evaluar críticamente contenido en plataformas digitales, facilitando acceso equitativo a mecanismos de verificación.

8. Metodología

8.1. Estrategia de investigación

Este proyecto adopta una estrategia de **Design and Creation** [16], apropiada para investigación en sistemas de información donde se desarrolla un artefacto tecnológico nuevo cuya construcción y evaluación constituyen la contribución principal al conocimiento. Según Hevner et al. [8], citados por Oates, esta estrategia requiere: (1) crear un artefacto viable, (2) abordar un problema relevante, (3) realizar evaluación rigurosa, (4) realizar contribuciones claras, (5) aplicar rigor en la construcción, (6) diseñar como proceso de búsqueda, y (7) comunicar efectivamente los resultados.

Técnicas de recolección de datos para evaluación:
Cuantitativas:

- Medición automatizada de métricas del sistema: tiempo de procesamiento, coste (tokens × precio), número de fuentes recuperadas, distribución de stances
- Evaluación mediante LLM-as-a-judge para valorar calidad de claims extraídos y veredictos generados
- Métricas de coordinación multiagente: task completion rate, latencia por agente

Cualitativas:

- Revisión manual de outputs del sistema (relevancia de claims, coherencia de veredictos)
- Estudios de caso detallados en diferentes temáticas

Herramientas: Python 3.12, Pydantic AI, FastAPI, React, OpenAI, Ollama, Selenium, etc.

8.2. Metodología de trabajo y desarrollo

Posibles estrategias para llevar a cabo el trabajo:

1. **Adaptar soluciones existentes:** Partir de sistemas académicos de fact-checking y extenderlos para contenido audiovisual
2. **Desarrollo de producto nuevo end-to-end:** Crear un sistema completo desde cero con interfaz web funcional
3. **Prototipo de investigación:** Implementar únicamente el pipeline de verificación sin interfaz de usuario
4. **Integración de servicios externos:** Combinar APIs de terceros existentes (Google Fact Check, transcripción comercial)

Estrategia escogida: Desarrollo de un producto nuevo end-to-end con interfaz web funcional (opción 2).

Justificación: Esta estrategia es la más apropiada porque permite evaluar la viabilidad práctica completa del sistema en condiciones reales de uso, no solo como concepto teórico. El desarrollo modular desde cero facilita la experimentación con diferentes configuraciones de LLMs y garantiza control total sobre el pipeline de verificación. La inclusión de interfaz web demuestra aplicabilidad para usuarios finales y permite recoger feedback sobre usabilidad real del sistema.

Proceso de desarrollo iterativo:

1. Implementación de componentes individuales con testing aislado
2. Integración end-to-end del pipeline (MVP funcional)
3. Refinamiento iterativo módulo por módulo basado en resultados
4. Evaluación comparativa con diferentes configuraciones de LLMs

Recursos técnicos: Prompting avanzado (chain-of-thought, few-shot learning), structured outputs con Pydantic, evaluación mediante LLM-as-a-judge, web scraping con Selenium, streaming en tiempo real (SSE). Conjunto de videos de evaluación en temáticas diversas.

9. Planificación del proyecto

9.1. Cronograma general

El proyecto se desarrolla durante el semestre académico 2025-2026, alineado con las entregas establecidas (Cuadro 1).

Hito	Entregable	Fecha límite
M1	Definición del TFM	12 oct 2025
M2	Estado del arte	2 nov 2025
M3	Implementación	14 dic 2025
M4	Redacción memoria preliminar	21 dic 2025
M4	Redacción memoria final	28 dic 2025
M4	Presentación audiovisual	6 ene 2026
M5	Documentación al tribunal	9 ene 2026
M5	Defensa pública	30 ene 2026

Cuadro 1: Cronograma general del proyecto

9.2. Descripción de fases

M1: Definición del TFM (hasta 12 oct 2025)

- Revisión bibliográfica sobre fact-checking automatizado y sistemas multiagente con LLMs
- Diseño de arquitectura del sistema y definición de componentes
- Selección de stack tecnológico
- Elaboración de propuesta de TFM

M2: Estado del arte (hasta 2 nov 2025)

- Análisis en profundidad de soluciones existentes
- Revisión de técnicas de evaluación de LLMs
- Setup inicial del proyecto y estructura de repositorio
- Documentación del estado del arte

M3: Implementación (hasta 14 dic 2025)

- Implementación de los cinco componentes del pipeline multiagente
- Integración end-to-end del sistema completo
- Desarrollo de API REST con streaming SSE e interfaz web
- Framework de evaluación de agentes
- MVP funcional operativo

M4: Evaluación y redacción (hasta 28 dic 2025)

- Experimentación iterativa para optimización de componentes

- Testing en conjunto de videos diversos
- Análisis de métricas y resultados
- Redacción de memoria del TFM
- Preparación de presentación audiovisual

M5: Defensa (hasta 30 ene 2026)

- Entrega de documentación final al tribunal
- Preparación de defensa pública
- Defensa del TFM

10. Bibliografía preliminar

Referencias

[1] Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. FEVEROUS: Fact extraction and verification over unstructured and structured information. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS 2021)*, volume 1, 2021.

[2] Yochai Benkler, Robert Faris, and Hal Roberts. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press, New York, 2018. ISBN 978-0-190-92303-4.

[3] Varich Boonsanong, Vidhisha Balachandran, Xiaochuang Han, Shangbin Feng, Lucy Lu Wang, and Yulia Tsvetkov. FACTS&EVIDENCE: An interactive tool for transparent fine-grained factual verification of machine-generated text. *arXiv preprint arXiv:2503.14797*, 2025. Available at https://arxiv.org/abs/2503.14797.

[4] Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 16477–16508. Association for Computational Linguistics, 2023. Available at https://aclanthology.org/2023.emnlp-main.557.pdf.

[5] Google for Developers. Google Fact Check Tools API. https://developers.google.com/fact-check/tools/api, 2024. Fact Check Explorer: https://toolbox.google.com/factcheck/explorer. Accessed: 2025-10-12.

[6] Naeemul Hassan, Chengkai Li, and Mark Tremayne. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1835–1838, Melbourne, Australia, 2015. ACM.

[7] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. ClaimBuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948, 2021. Available at https://arxiv.org/abs/2112.01488.

[8] Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS Quarterly*, 28(1):75–105, 2004.

[9] Chip Huyen. *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*. O'Reilly Media, Sebastopol, CA, 2022. ISBN 978-1-098-10796-3.

[10] HySonLab. FactAgent github repository. https://github.com/HySonLab/FactAgent, 2025. Accessed: 2025-10-12.

[11] Kalmi LLC. Fact-Checker implementation. https://github.com/kalmiallc/fact-checker, 2024. Accessed: 2025-10-12.

[12] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 9459–9474, 2020.

[13] Potsawee Manakul, Adian Liusie, and Mark JF Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9004–9017, 2023.

[14] Preslav Nakov, Firoj Alam, Alberto Barrón-Cedeño, Giovanni Da San Martino, Maram Gupta, Fatima Haouari, Maram Hasanain, Watheq Husin, Georgi Karadzhov, Maria Pontiki, et al. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual

setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361. Association for Computational Linguistics, 2023.

[15] Hieu Nguyen, Minh Pham, and Quang Tran. FactAgent: Towards robust fact-checking with multi-agent systems and advanced evidence retrieval. *arXiv preprint arXiv:2506.17878*, 2025. Preprint. Available at https://arxiv.org/pdf/2506.17878.

[16] Briony J. Oates. *Researching Information Systems and Computing*. SAGE Publications Ltd., London, 2006. ISBN 9781446203620.

[17] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024. Updated version.

[18] OpenAI. Latency optimization for LLM applications. https://platform.openai.com/docs/guides/latency-optimization, 2024. Accessed: 2025-10-12.

[19] Pydantic Team. Pydantic documentation: Data validation using python type hints. https://docs.pydantic.dev/, 2024. Accessed: 2025-10-01.

[20] Sebastián Ramírez. FastAPI: Modern, fast (high-performance) web framework for building APIs. https://fastapi.tiangolo.com/, 2024. Accessed: 2025-10-01.

[21] Sebastian Raschka. LLM evaluation: 4 practical approaches. https://sebastianraschka.com/blog/2025/llm-evaluation-4-approaches.html, 2025. Accessed: 2025-10-12.

[22] Sebastian Ruder. The evolving landscape of LLM evaluation. https://www.ruder.io/the-evolving-landscape-of-llm-evaluation/, 2025. Accessed: 2025-10-12.

[23] Shaden Shaar, Alberto Barrón-Cedeño, Giovanni Da San Martino, and Preslav Nakov. Overview of the CLEF-2023 CheckThat! lab task 1 on check-worthiness estimation in multimodal and multigenre content. In *Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023)*. CEUR-WS.org, 2023. Available at https://aclanthology.org/2023.semeval-1.311.pdf.

[24] The Turing Post. Agents recap: State of AI agents in 2024. https://www.turingpost.com/p/agentsrecap, 2024. Accessed: 2025-10-12.

[25] The Turing Post. Profiling large language models: Understanding performance characteristics. https://www.turingpost.com/p/profiling, 2024. Accessed: 2025-10-12.

[26] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 809–819, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[27] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[28] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550. Association for Computational Linguistics, 2020.

[29] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345, 2024.

[30] Claire Wardle and Hossein Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policy making. Report DGI(2017)09, Council of Europe, Strasbourg, France, 2017.

[31] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

[32] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

[33] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.