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

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.

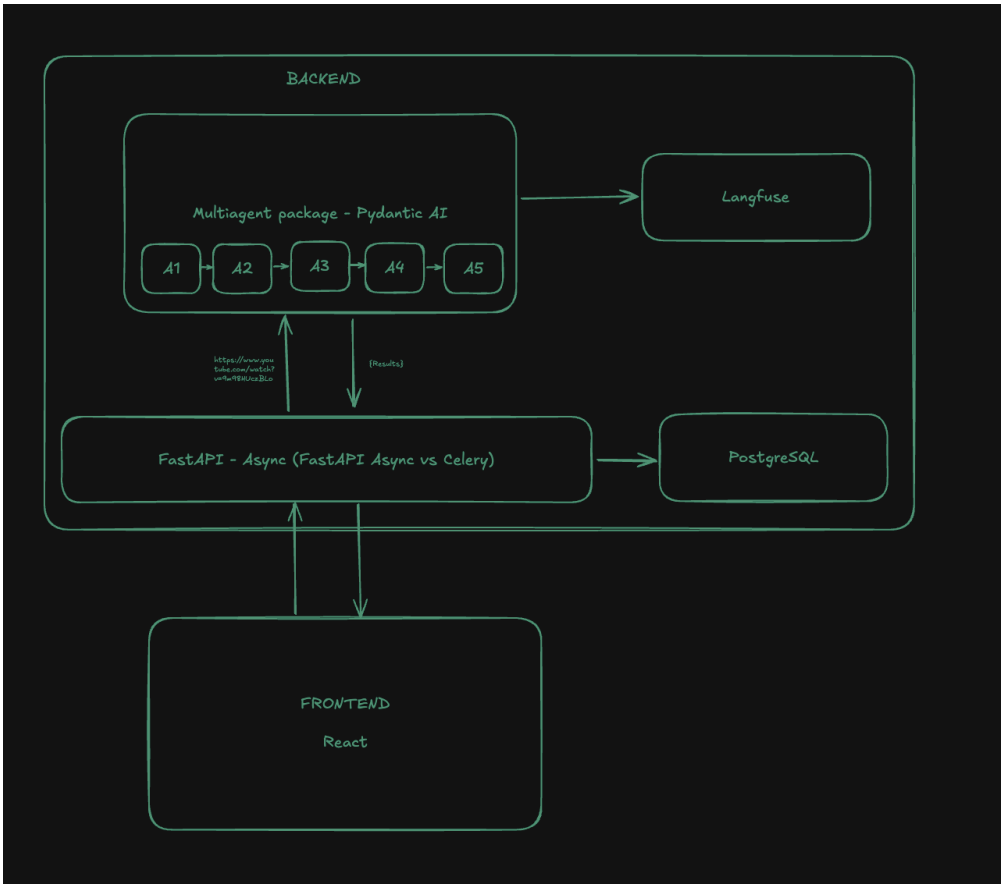

Figura 1: Arquitectura del sistema multiagente propuesto. Pipeline de cinco componentes especializados procesando secuencialmente desde transcripción hasta veredicto final.

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, 28]. 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 [24], FEVEROUS [1] y trabajos recientes sobre verificación científica [26] 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. **Sistema completo operativo:** A diferencia de prototipos académicos, se desarrollará una aplicación funcional con interfaz web.
2. **Análisis comparativo de LLMs:** Evaluación empírica de modelos comerciales (GPT-4o-mini) versus open-source (Ollama Qwen 3), documentando compromisos entre coste, rendimiento y calidad.
3. **Metodología de evaluación:** Protocolo de evaluación cualitativa y cuantitativa aplicable a sistemas similares.
4. **Análisis de implicaciones éticas:** Estudio de sesgos, limitaciones y riesgos de la automatización del fact-checking.

4.5. Alcance del proyecto

Incluye:

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

No incluye:

- Generación propia de transcripciones desde audio (speech-to-text)
- Verificación de contenido visual (detección de deepfakes, manipulación de imágenes)
- Sistema de caché o base de datos persistente de verificaciones
- Verificación en tiempo real o procesamiento batch a gran escala
- Contenido de plataformas distintas a YouTube

5. Motivación personal

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

Experiencia profesional: Como Machine Learning Engineer con más de 5 años de experiencia en Data Science, he trabajado en proyectos que combinan 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 la 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 transitar desde roles puramente técnicos hacia posiciones que combinan 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

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)

OE1. Diseño de arquitectura multiagente

- Definir pipeline de procesamiento con componentes especializados
- Establecer esquemas de datos estructurados para comunicación inter-agente
- Diseñar sistema de coordinación y flujo de información

OE2. Implementación de componentes del sistema

- Desarrollar módulo de extracción de transcripciones desde YouTube
- Implementar agente de identificación de claims con clasificación de relevancia
- Crear generador de consultas de búsqueda optimizadas
- Desarrollar sistema de web scraping con evaluación de fiabilidad de fuentes
- Implementar generador de veredictos con razonamiento explicable

OE3. Optimización de uso de LLMs

- Comparar rendimiento entre modelos comerciales (GPT-4o-mini) y open-source (Ollama Qwen 3)
- Analizar compromisos entre coste, latencia y calidad de outputs
- Evaluar estrategias de optimización de prompts y structured outputs

OE4. Evaluación del sistema

- Definir métricas de rendimiento técnico (tiempo, coste, throughput)
- Establecer protocolos de evaluación cualitativa de veredictos
- Realizar estudios de caso en diferentes temáticas
- Analizar limitaciones y casos de fallo del sistema

OE5. Implementación de sistema end-to-end

- Desarrollar API REST con streaming en tiempo real

- Crear interfaz web intuitiva para usuarios finales
- Garantizar reproducibilidad y documentación exhaustiva

OE6. Análisis de implicaciones éticas

- Identificar sesgos potenciales en selección de fuentes y generación de veredictos
- Analizar riesgos de automatización de fact-checking
- Proponer salvaguardas y mejores prácticas

7. Competencia de Compromiso Ético y Global (CCEG)

Esta sección evalúa el impacto del proyecto en las tres dimensiones de la Competencia de Compromiso Ético y Global establecida por la UOC.

7.1. Sostenibilidad y Objetivos de Desarrollo Sostenible (ODS)

El uso de modelos de lenguaje de gran escala implica un consumo energético significativo. Este proyecto aborda explícitamente esta preocupación mediante:

- **Evaluación de alternativas eficientes:** Comparación entre modelos comerciales cloud y modelos open-source ejecutables localmente, priorizando modelos locales durante experimentación.
- **Análisis de compromisos:** Documentación transparente del consumo de recursos (tokens procesados, tiempo de ejecución, coste económico).
- **Optimización:** Refinamiento de prompts para minimizar tokens procesados sin sacrificar calidad.

Relación con ODS:

- **ODS 4 (Educación de Calidad):** Herramienta de alfabetización mediática que promueve pensamiento crítico.
- **ODS 16 (Paz, Justicia e Instituciones Sólidas):** Combate la desinformación que erosiona instituciones democráticas.
- **ODS 10 (Reducción de Desigualdades):** Democratiza acceso a verificación de hechos.

En las conclusiones se incluirá análisis crítico sobre si la automatización justifica el consumo energético frente a beneficios sociales.

7.2. Comportamiento ético y responsabilidad social

Impactos éticos identificados y estrategias de mitigación:

(1) Automatización y empleo: El sistema podría percibirse como amenaza para verificadores profesionales. *Estrategia:* Posicionamiento como herramienta de apoyo que automatiza pasos preliminares, permitiendo a profesionales enfocarse en análisis complejos que requieren juicio humano.

(2) Sesgo algorítmico: Los LLMs pueden perpetuar sesgos presentes en datos de entrenamiento. *Mitigación:* Transparencia total en fuentes consultadas, presentación de evidencias con diferentes posturas, evaluación explícita de confianza en veredictos, y advertencias claras sobre limitaciones.

(3) Riesgo de amplificación: Al citar afirmaciones incorrectas durante verificación, existe riesgo de amplificar desinformación. *Mitigación:* Presentación siempre contextualizada con veredicto, no indexación pública de claims individuales sin contexto, enfoque en proceso educativo más que en listados de falsedades.

(4) Privacidad: El proyecto respeta políticas de YouTube y no procesa contenido privado ni viola términos de uso de plataformas.

Como profesional de Data Science, este proyecto se adhiere a principios deontológicos de transparencia, no maleficencia, beneficencia social y responsabilidad ante consecuencias imprevistas.

7.3. Diversidad, género y derechos humanos

Accesibilidad: La interfaz web se diseñará siguiendo principios WCAG 2.1: contraste de colores adecuado, navegación por teclado, textos alternativos para elementos visuales, diseño responsive adaptable a diferentes dispositivos.

Diversidad lingüística: Se reconoce que los modelos LLM tienen mejor rendimiento en inglés. Mitigación mediante soporte experimental para español y documentación explícita de limitaciones lingüísticas.

Diversidad de perspectivas: Las fuentes web indexadas pueden tener sesgo geográfico/cultural (predominancia de contenido occidental). El sistema evaluará diversidad de perspectivas cuando estén disponibles y documentará esta limitación.

Derechos fundamentales:

- **Derecho a la información:** Impacto positivo al facilitar acceso a verificación.
- **Libertad de expresión:** El sistema no censura contenido, solo proporciona contexto adicional.
- **Privacidad:** No procesa datos personales ni información identificable de usuarios.

8. Metodología

8.1. Estrategia de investigación

Este proyecto adopta una metodología de **Investigación y Desarrollo (Design Science Research)**, combinando investigación en ciencia de datos con ingeniería de software aplicada.

La estrategia seleccionada es el **desarrollo end-to-end con evaluación comparativa**, que combina implementación completa del pipeline (valor ingenieril), experimentación con diferentes modelos (valor investigativo), evaluación cuantitativa y cualitativa (rigor científico), e interfaz funcional (aplicabilidad real). Esta aproximación es apropiada porque:

- El problema es complejo y multidisciplinar, requiriendo integración de NLP, recuperación de información y diseño de sistemas
- El resultado debe ser un sistema operativo funcional que aporte valor práctico, no solo análisis teórico
- Las decisiones técnicas se validan mediante experimentación empírica con diferentes configuraciones
- El desarrollo modular permite mejoras incrementales basadas en evaluación continua

8.2. Metodología de desarrollo

El sistema sigue una **arquitectura de pipeline** (Figura 1) con cinco agentes especializados:

1. **Transcriptor:** Extrae texto desde videos de YouTube usando APIs oficiales
2. **Claim Extractor:** Identifica afirmaciones factuales verificables mediante LLMs
3. **Query Generator:** Crea consultas de búsqueda optimizadas para motores de búsqueda
4. **Online Search:** Recupera y evalúa evidencias web con análisis de credibilidad de fuentes
5. **Output Generator:** Sintetiza veredictos fundamentados con razonamiento explicable

Cada componente opera sobre esquemas de datos estructurados definidos con Pydantic, garantizando validación de tipos y contratos claros entre módulos.

Proceso de desarrollo iterativo:

1. Implementación básica de cada componente individualmente
2. Integración E2E del pipeline completo (MVP funcional)
3. Refinamiento iterativo módulo por módulo
4. Evaluación comparativa con diferentes configuraciones de LLMs

8.3. Técnicas y herramientas

Técnicas de Procesamiento de Lenguaje Natural:

- Prompting avanzado con LLMs (chain-of-thought, few-shot learning)
- Structured outputs mediante Pydantic AI para garantizar formato de respuestas
- Análisis de stance (postura de evidencias: apoya/refuta/neutral/no clara)

Técnicas de Recuperación de Información:

- Web scraping con Selenium para extracción de contenido
- Evaluación de credibilidad mediante análisis WHOIS y características de sitio
- Estrategias de búsqueda diversificadas (múltiples consultas por claim)

Stack Tecnológico:

- **Backend:** Python 3.12, Pydantic AI, FastAPI, Uvicorn
- **LLMs:** OpenAI GPT-4o-mini (comercial), Ollama con Qwen 3 (open-source local)
- **Frontend:** React 18.3, Vite, Tailwind CSS
- **Herramientas de desarrollo:** uv (gestión de dependencias), ruff (linting), mypy (type checking)
- **Control de versiones:** Git, GitHub

8.4. Métodos de evaluación

Métricas cuantitativas:

- Tiempo de procesamiento por video
- Coste por verificación (tokens consumidos × precio por token)
- Número de fuentes recuperadas por claim
- Distribución de stances (evidencias a favor/contra/neutral)

Evaluación cualitativa:

- Revisión manual de relevancia de claims extraídos
- Análisis de coherencia y fundamentación de veredictos
- Comparación con fact-checks profesionales cuando disponibles
- Identificación de casos de fallo y limitaciones

Conjunto de evaluación: Dataset de 15-20 videos diversos cubriendo diferentes temáticas (economía, política, ciencia, pseudociencia) y niveles de controversia.

9. Planificación del proyecto

9.1. Cronograma general

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

Hito	Entregable	Fecha límite
M1	Definición del TFM	12 oct 2025
M1	Comité ética y confidencialidad	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 (presente documento)

M2: Estado del arte (hasta 2 nov 2025)

- Análisis en profundidad de soluciones existentes (FEVER, FEVEROUS, FactAgent)
- 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
- Creación de interfaz web en React
- MVP funcional operativo

M4: Evaluación y redacción (hasta 28 dic 2025)

- Evaluación comparativa de LLMs (GPT-4o-mini vs Ollama Qwen 3)
- 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] 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.

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

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

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

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

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

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

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

[21] 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.

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

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.