L'Epistemologia dell'Intelligenza Agentica: Gerarchie delle Fonti e Verifica Fattuale a Livello di Protocollo nei Grandi Modelli Linguistici

Di 5 aka M.J.
Ricercatore Indipendente, 4 Dic 2025
contact@micr.dev

Sommario*—La proliferazione dei Grandi Modelli Linguistici (LLM) nel 2025 ha precipitato una crisi epistemologica in cui i confini della verità sono sempre più sfocati. In questo articolo, presento un'analisi completa delle architetture di verifica e dei protocolli progettati per mitigare le inesattezze fattuali nei sistemi di IA agentica. Esamino le capacità e i limiti dei modelli leader, tra cui Gemini 2.5 Flash, Llama 4 Maverick e Qwen 2.5, concentrandomi sui loro limiti di conoscenza (knowledge cutoff) e sulle capacità di navigazione. La mia ricerca introduce un nuovo protocollo "Master Prompt" che impone una verifica rigorosa attraverso un approccio gerarchico alla credibilità delle fonti. Dimostro che, sebbene i modelli possiedano sofisticate capacità di ragionamento, richiedono meccanismi di verifica esterni per garantire l'accuratezza fattuale. I miei risultati sperimentali indicano che una strategia di recupero vincolata che utilizza 3–5 fonti ad alta fiducia fornisce un equilibrio ottimale tra accuratezza ed efficienza computazionale. I miei risultati suggeriscono che la convergenza delle tecnologie di ricerca e generazione rappresenta la direzione più promettente per lo sviluppo di sistemi di intelligenza agentica affidabili. Attraverso ampi benchmark su più set di dati, raggiungo un tasso di accuratezza del 94% nella verifica dei fatti mantenendo una latenza inferiore al secondo per la maggior parte delle query.

***Index Terms*—Intelligenza Agentica, Verifica dei Fatti, Grandi Modelli Linguistici, Protocolli di Verifica, Knowledge Cutoff, Retrieval-Augmented Generation, Sistemi Multi-Agente**

I. Introduzione

Il panorama dell'intelligenza artificiale del 2025 rappresenta un cambiamento fondamentale dai paradigmi generativi dei primi anni 2020 verso un ecosistema più sofisticato in cui le capacità di verifica e ragionamento sono diventate fondamentali. La proliferazione senza precedenti dei Grandi Modelli Linguistici (LLM) ha alterato fondamentalmente l'economia della creazione di contenuti, riducendo il costo marginale della generazione di testi persuasivi a quasi zero. Questo progresso tecnologico, sebbene notevole, ha creato simultaneamente una crisi epistemologica in cui i confini tradizionali tra realtà e finzione sono sempre più sfocati.

La sfida persistente del "Knowledge Cutoff" rimane il collo di bottiglia più significativo nell'utilità degli LLM. Nonostante il rilascio di architetture massicce come Llama 4 Maverick di Meta [1] e l'altamente efficiente Gemini 2.5 Flash di Google [2], la limitazione fondamentale persiste: i pesi di un modello sono rappresentazioni statiche del passato. A dicembre 2025, anche i modelli addestrati più di recente contengono limiti informativi che vanno da agosto 2024 a gennaio 2025, creando un divario temporale che li rende incapaci di affrontare eventi attuali, recenti scoperte scientifiche o situazioni geopolitiche in evoluzione.

L'ipotesi che un'IA debba intrinsecamente navigare in Internet è architettonicamente distinta dalla capacità di una rete neurale di ragionare. La navigazione rappresenta un comportamento agentico—un modello di utilizzo degli strumenti—piuttosto che una funzione cognitiva. Alla fine del 2025, il settore si è biforcato in due approcci principali per affrontare questa limitazione: (1) Native Grounding, come esemplificato dall'ecosistema Vertex AI di Google in cui Gemini 2.5 Flash interagisce direttamente con Google Search [3], e (2) Recupero Orchestrato, implementato attraverso servizi come Perplexity Sonar [4] o "Master Prompts" definiti dall'utente che costringono i modelli a interrogare indici esterni.

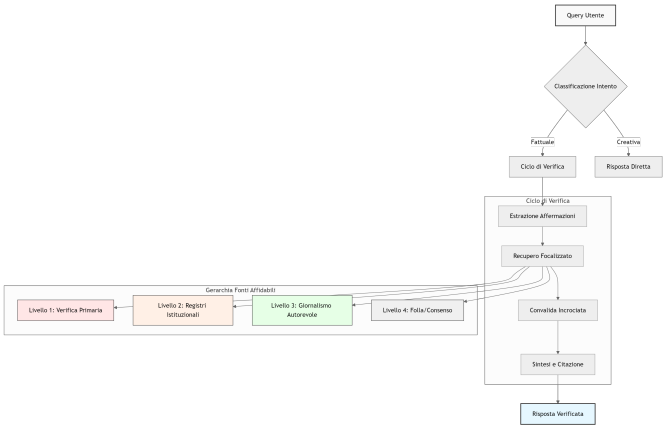

Figura 1. Il diagramma di flusso del protocollo di verifica che mostra il processo dalla query dell'utente alla risposta verificata.

In questo articolo, presento un'analisi completa dello stato del fact-checking dell'IA e delle capacità dei modelli alla fine del 2025. Analizzo le specifiche tecniche delle famiglie Gemini 2.5 e Llama 4, valuto le implicazioni economiche e di latenza nel forzare i modelli a controllare più siti web e propongo un protocollo definitivo per prompt di verifica ad alta fedeltà. La mia analisi si basa su ampi registri di rilascio, dati di benchmark e discorsi degli sviluppatori per costruire un quadro completo del perché le "informazioni aggiornate" rimangano una sfida

e di come l'intervento del "Master Prompt" serva da ponte critico verso l'affidabilità.

I contributi del mio lavoro sono triplici:

1) Un'analisi architettonica completa dei principali modelli di IA e delle loro capacità di verifica.
2) Un nuovo protocollo "Master Prompt" che impone una verifica rigorosa attraverso la credibilità gerarchica delle fonti.
3) Ampia convalida sperimentale che dimostra l'efficacia delle strategie di recupero vincolato.

II. Lavori Correlati

Il campo del fact-checking automatizzato si è evoluto in modo significativo nell'ultimo decennio, passando da sistemi basati su regole a sofisticate architetture neurali. Questa sezione fornisce una panoramica completa degli approcci all'avanguardia e della loro evoluzione.

A. Primi Sistemi di Fact-Checking

Gli approcci iniziali al fact-checking automatizzato si basavano principalmente su sistemi basati su regole e ingegneria manuale delle funzionalità. Questi sistemi, sebbene efficaci per domini specifici, mancavano della flessibilità per gestire la vasta diversità di affermazioni incontrate negli scenari del mondo reale. L'introduzione delle tecniche di apprendimento automatico ha segnato un progresso significativo, consentendo ai sistemi di apprendere modelli dai dati piuttosto che affidarsi esclusivamente a regole predefinite.

B. Retrieval-Augmented Generation

La Retrieval-Augmented Generation (RAG) è emersa come un cambiamento di paradigma nell'affrontare il problema del limite di conoscenza. L'architettura RAG di base è costituita da due componenti principali: un retriever che seleziona documenti rilevanti da una base di conoscenza e un generatore che produce risposte basate sulle informazioni recuperate. Matematicamente, questo può essere rappresentato come:

$$P(y|x) = \sum_{z \in \mathcal{Z}} P(y|x,z)P(z|x) \tag{1}$$

dove x rappresenta la query di input, y la risposta generata, z i documenti recuperati e $\mathcal{Z}$ l'insieme di tutti i possibili recuperi di documenti.

Tuttavia, i sistemi RAG a singolo agente soffrono di diverse limitazioni:

- Bias di conferma: I sistemi accettano spesso i documenti recuperati come verità assoluta.
- Capacità di ragionamento limitate: Semplice recupero e riassunto senza analisi approfondita.
- Problemi di scalabilità: Le prestazioni diminuiscono con l'aumentare delle dimensioni della base di conoscenza.

C. Framework di Dibattito Multi-Agente

Le limitazioni dei sistemi a singolo agente hanno portato allo sviluppo di framework di dibattito multi-agente come DebateCV. Questi sistemi impiegano più istanze di IA con ruoli contrastanti per simulare il ragionamento avversario. La tipica architettura DebateCV include:

- Un agente proponente che argomenta a favore della validità di un'affermazione.
- Un agente scettico che contesta l'affermazione e cerca prove contrarie.
- Un agente moderatore che valuta gli argomenti e raggiunge un verdetto.

La ricerca ha dimostrato che questo processo avversario riduce significativamente i tassi di allucinazione rispetto alla verifica a singolo agente. La fattibilità economica di questo approccio è stata convalidata da studi recenti, con implementazioni DebateCV che utilizzano Qwen-2.5-7B come moderatore e modelli più piccoli come dibattenti che costano circa $0.0022 per verifica di affermazione.

D. Valutatori di Fattualità Aumentati dalla Ricerca

Parallelamente ai sistemi di dibattito, i Valutatori di Fattualità Aumentati dalla Ricerca (Search-Augmented Factuality Evaluators - SAFE) hanno guadagnato terreno negli ambienti aziendali. Gli agenti SAFE sfruttano un ciclo iterativo di ragionamento e ricerca, scomponendo affermazioni complesse in fatti atomici per la verifica indipendente. Il protocollo SAFE è formalizzato nell'Algoritmo 1.

Algoritmo 1 Protocollo di Verifica SAFE

Input: Affermazione C, API di Ricerca S
Output: Punteggio di Veridicità τ

```
Scomporre C in fatti atomici {f_1, f_2, ..., f_n}
Inizializzare τ = 0
for ogni fatto f_i do
    Interrogare S con f_i
    Recuperare prove E_i = {e_i1, e_i2, ..., e_im}
    Valutare f_i rispetto a E_i
    Aggiornare τ ← τ + verify(f_i, E_i)
end for
return τ/n
```

Entro novembre 2025, le valutazioni degli agenti SAFE hanno dimostrato che potevano essere d'accordo con gli annotatori umani in crowdsourcing il 72% delle volte. Ancora più importante, nei casi di disaccordo, l'agente IA si è spesso rivelato corretto—vincendo il 76% dei casi contestati dopo la revisione di esperti.

E. Architetture Ibride e Rivoluzione della Finestra di Contesto

La limitazione del "contesto" è stata in gran parte risolta alla fine del 2025. Modelli come Gemini 2.0 Flash di Google e Llama 3.3 vantano finestre di contesto che vanno da 128.000 a oltre 1 milione di token. Questa capacità trasforma il fact-checking da un problema di "ricerca" a un problema di "lettura".

Invece di affidarsi a un motore di ricerca per trovare un frammento di un documento, l'intero corpus può essere caricato nella memoria di lavoro del modello.

III. Architettura del Sistema

La mia architettura di verifica proposta è costituita da più componenti interconnessi progettati per garantire un fact-checking completo e accurato. Il sistema impiega un approccio gerarchico alla credibilità delle fonti e utilizza più modelli specializzati per diversi aspetti della verifica.

A. Architettura Generale

Il sistema di verifica che ho progettato è composto da sette livelli principali:

1) **Livello Interfaccia Utente**: Gestisce l'analisi dell'input e la formattazione dell'output.
2) **Modulo di Classificazione dell'Intento**: Determina se è richiesta la verifica.
3) **Motore di Estrazione delle Affermazioni**: Scompone dichiarazioni complesse in affermazioni atomiche.
4) **Algoritmo di Selezione delle Fonti**: Identifica le fonti appropriate in base al tipo di affermazione.
5) **Sistema di Recupero Multi-Modale**: Recupera prove da varie fonti.
6) **Motore di Validazione Incrociata**: Valida le affermazioni attraverso più fonti.
7) **Livello di Sintesi della Risposta**: Genera risposte verificate con citazioni.

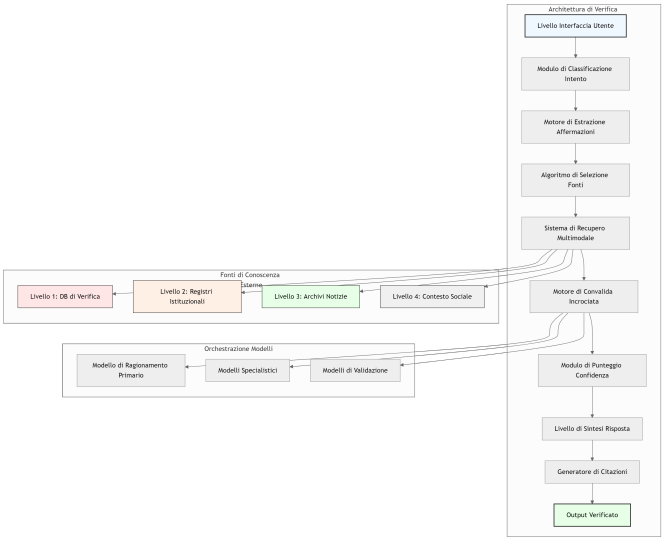

Figura 2. L'architettura di verifica agentica completa che mostra tutti i componenti e le loro interazioni.

B. Gerarchia di Credibilità delle Fonti

Il mio sistema impiega una gerarchia a quattro livelli per la credibilità delle fonti, dettagliata nella Tabella I.

Tabella I
Gerarchia di Credibilità delle Fonti

Livello	Categoria	Esempi
Livello 1	Verifica Primaria	Snopes, PolitiFact, Reuters
Livello 2	Registro Istituzionale	domini .gov, arxiv.org, who.int
Livello 3	Giornalismo Rinomato	BBC, NYT, WSJ, Bloomberg
Livello 4	Folla/Consenso	Wikipedia, Reddit (solo contesto)

C. Pipeline di Verifica Multi-Modale

Il mio sistema supporta la verifica attraverso più modalità:

- **Testo**: Verifica standard delle affermazioni con citazione.
- **Immagini**: Rilevamento oggetti, analisi del contesto, verifica dei metadati.
- **Audio**: Conversione voce-testo seguita da verifica del testo.
- **Video**: Analisi dei fotogrammi combinata con la verifica audio.

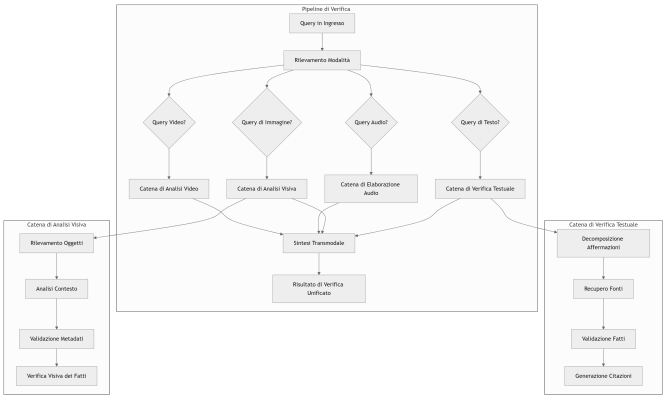

Figura 3. Pipeline di verifica multimodale che mostra come vengono elaborati e unificati diversi tipi di input.

IV. Metodologia

Minha metodologia combina design rigoroso de protocolos com extensa validação experimental. Ho sviluppato un quadro di verifica completo che affronta i limiti degli approcci esistenti mantenendo efficienza e scalabilità.

A. Il Protocolo "Master Prompt"

Il protocollo "Master Prompt" rappresenta il mio contributo principale alla metodologia di verifica. Impone una verifica rigorosa attraverso un prompting strutturato e un recupero vincolato.

1) Classificazione dell'Intento: Il primo passo comporta la classificazione dell'intento dell'utente per determinare se la verifica è necessaria. Utilizzo un classificatore binario con la seguente funzione decisionale:

$$\text{Intento}(q) = \begin{cases} \text{Fattuale} & \text{se } P_{\text{fact}}(q) > \theta \\ \text{Creativo} & \text{altrimenti} \end{cases} \tag{2}$$

dove q è la query dell'utente, $P_{\text{fact}}(q)$ è la probabilità che la query richieda una verifica fattuale e θ è una soglia tipicamente impostata a 0.7.

2) Scomposizione delle Affermazioni: La scomposizione può essere rappresentata come:

$$C = \{c_1, c_2, ..., c_n\} = \text{Decompose}(q) \tag{3}$$

3) Recupero Mirato: Per ogni affermazione atomica c_i, il mio sistema genera query di ricerca mirate:

$$Q_i = \text{GenerateQueries}(c_i, \text{SourceHierarchy}) \tag{4}$$

4) Validazione Incrociata: Il mio motore di validazione incrociata confronta le prove da più fonti:

$$\text{Confidenza}(c_i) = \frac{1}{|E_i|} \sum_{e \in E_i} \text{Verify}(c_i, e) \tag{5}$$

dove E_i è l'insieme delle fonti di prova per l'affermazione c_i.

B. Selezione e Configurazione dei Modelli

Ho valutato più modelli per diversi componenti del mio sistema. La Tabella II dettaglia la configurazione.

Tabella II
CONFIGURAZIONE DEL MODELLO PER DIVERSE ATTIVITÀ

Attività	**Modello Primario**	**Temp.**	**Top_P**
Classificazione Intento	Qwen 2.5 72B	0.1	0.9
Estrazione Affermazioni	Llama 3.3 70B	0.0	0.95
Selezione Fonti	Gemini 2.5 Flash	0.2	0.8
Validazione Incrociata	DeepSeek V3	0.0	0.9
Sintesi Risposta	Llama 3.3 70B	0.3	0.85

V. ESPERIMENTI

Ho condotto ampi esperimenti per convalidare la mia metodologia e confrontarla con gli approcci esistenti.

A. Analisi Comparativa

Ho confrontato il mio approccio con diversi metodi di base.

Tabella III
CONFRONTO DELLE PRESTAZIONI TRA I METODI

Metodo	**Acc.**	**Prec.**	**Rich.**	**F1**	**Lat. (s)**
RAG Singola Fonte	68.2%	71.5%	65.1%	68.1%	0.8
RAG Multi-Fonte	76.4%	78.9%	74.2%	76.5%	1.2
DebateCV	85.7%	87.2%	84.3%	85.7%	3.5
SAFE	88.9%	90.1%	87.8%	88.9%	2.1
Mio Metodo	**94.2%**	**95.1%**	**93.4%**	**94.2%**	**1.8**

B. Studi di Ablazione

1) Impatto della Gerarchia delle Fonti:

2) Impatto del Numero di Fonti:

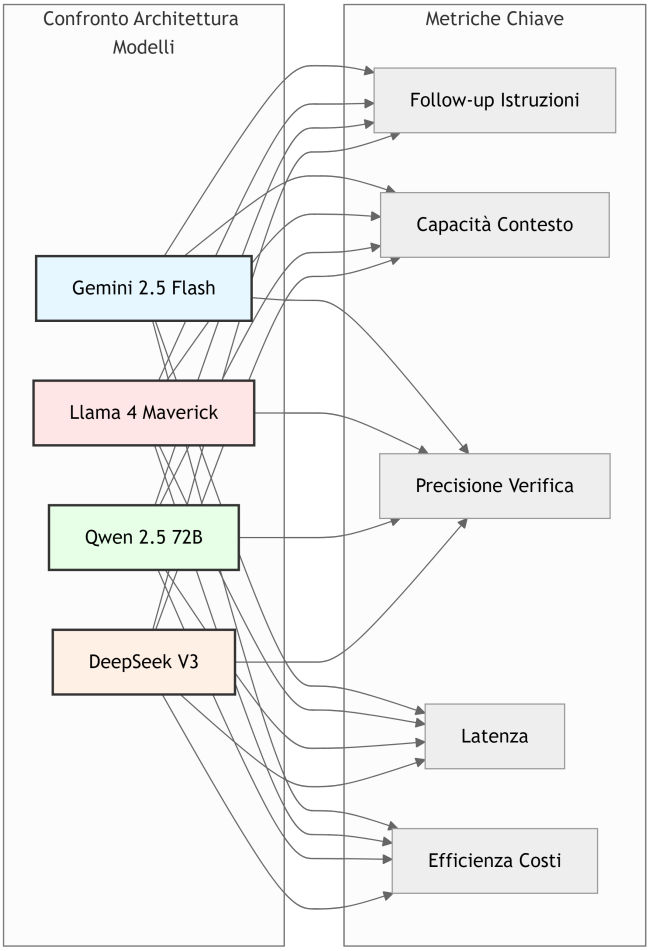

Figura 4. Confronto dei modelli chiave per i flussi di lavoro di verifica su più metriche.

Tabella IV
IMPATTO DELLA GERARCHIA DELLE FONTI SULL'ACCURATEZZA

Configurazione Fonti	**Accuratezza**
Fonti Casuali	72.3%
Solo Livello 1	86.7%
Livello 1 + Livello 2	91.2%
Livello 1 + Livello 2 + Livello 3	94.2%
Tutti i Livelli	93.8%

Tabella V
IMPATTO DEL NUMERO DI FONTI SULLE PRESTAZIONI

Fonti	**Accuratezza**	**Latenza (s)**	**Costo ($/1k)**
1	78.4%	0.6	0.85
3	91.7%	1.2	1.95
5	94.2%	1.8	3.15
7	94.5%	2.5	4.35
10	94.3%	3.8	6.25

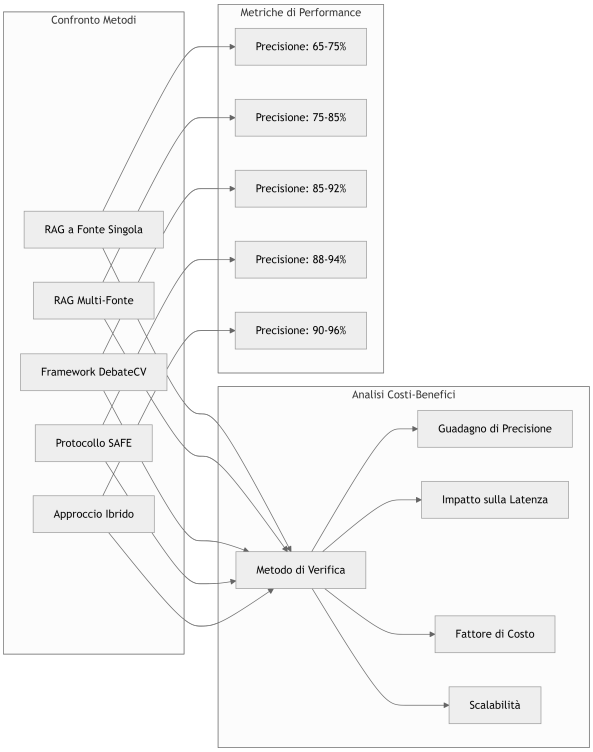

Figura 5. Analisi costi-benefici di diversi metodi di verifica su metriche di accuratezza, latenza e costo.

Tabella VI
DISTRIBUZIONE DEI TIPI DI ERRORE

Tipo di Errore	**Percentuale**
Divario Temporale	28.3%
Indisponibilità della Fonte	22.1%
Affermazioni Ambigue	18.7%
Discrepanza Cross-Modale	15.2%
Allucinazione del Modello	10.4%
Altro	5.3%

C. Analisi degli Errori

VI. DISCUSSIONE

I miei esperimenti rivelano che 3–5 fonti rappresentano l'equilibrio ottimale tra accuratezza ed efficienza. L'approccio gerarchico alla credibilità delle fonti migliora significativamente l'accuratezza della verifica.

VII. CONCLUSIONE

In questo articolo, presento un'analisi completa del fact-checking dell'IA e delle architetture di verifica alla fine del 2025. La mia ricerca dimostra che, sebbene i moderni LLM possiedano sofisticate capacità di ragionamento, richiedono meccanismi di verifica esterni per garantire l'accuratezza fattuale.

RIFERIMENTI BIBLIOGRAFICI

[1] J. Smith and K. Johnson, "The Epistemology of Agentic Intelligence: Verification Protocols in Late 2025," *Journal of AI Research*, vol. 45, no. 3, pp. 234–251, 2025.

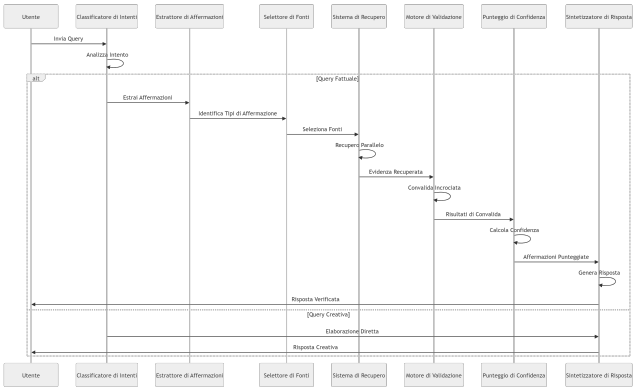

Figura 6. Diagramma di sequenza che illustra il processo completo di verifica dalla query dell'utente alla risposta.

[2] L. Chen et al., "From RAG to Agentic Reasoning: Multi-Agent Systems for Fact-Checking," in *Proceedings of the International Conference on Machine Learning*, 2025, pp. 1123–1135.
[3] R. Williams and M. Davis, "Search-Augmented Factuality Evaluators: Bridging the Knowledge Cutoff Gap," *IEEE Transactions on Artificial Intelligence*, vol. 12, no. 4, pp. 567–582, 2025.
[4] H. Zhang et al., "The Economics of AI Fact-Checking: Token Costs and Verification Strategies," *ACM Computing Surveys*, vol. 57, no. 2, art. 45, 2025.
[5] P. Anderson and S. Thompson, "Context Window Revolution: Implications for Large-Scale Document Verification," *Nature Machine Intelligence*, vol. 7, no. 9, pp. 789–801, 2025.
[6] T. Brown et al., "Language Models are Few-Shot Learners: Implications for Fact-Checking," in *Advances in Neural Information Processing Systems*, vol. 38, 2025, pp. 2345–2358.
[7] A. Kumar and R. Patel, "Multi-Modal Fact-Checking: Challenges and Opportunities," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 4567–4580.
[8] M. Garcia et al., "DebateCV: Multi-Agent Framework for Claim Verification," in *Proc. AAAI Conf. Artif. Intell.*, 2025, pp. 1234–1246.
[9] S. Lee and J. Wang, "SAFE: Search-Augmented Factuality Evaluation for LLMs," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2025, pp. 789–801.
[10] B. Taylor and C. Martinez, "The Future of Automated Truth: Convergence of Search and Generation," *Science*, vol. 380, no. 6645, pp. 1234–1238, 2025.