Claim & Core Insight	Quantitative Evidence & Context	Supporting Paper
Domain-Tuned SLMs Can Outperform Larger LLMs	• A fine-tuned BERT model achieved a +9.0% relative F1-score improvement over the best-performing GPT-3.5-turbo configuration (76.5% vs. 70.2%) on the GossipCop dataset \cite{hu2024bad} • A supervised RoBERTa-Base classifier outperformed zero-shot GPT-3.5-Turbo by +9.6 absolute points in F1-macro on the FA-KES dataset (52.9% vs. 43.3%) \cite{leite2023detecting} • A fine-tuned BioBERT for clinical claims achieved 80.2% accuracy on CliniFact, significantly outperforming both zero-shot (34.3%) and fine-tuned (53.6%) Llama3-70B \cite{zhang2025dataset}	3
Advanced Prompting is Better than Standard CoT	• The HiSS (Hierarchical Step-by-Step) method surpassed vanilla CoT by +9.5 absolute points in F1-score on the RAWFC dataset (53.9% vs. 44.4%). On the LIAR dataset,	1

	HiSS outperformed vanilla CoT by +7.1 points in F1-score (31.3% vs. 24.2%). HiSS also outperformed the more advanced ReAct agent framework by +4.1 absolute points on RAWFC, demonstrating the value of its hierarchical structure. \cite{zhang2023towards}	
RAG Provides Significant Performance Gains	• Providing external context (RAG) to GPT-4 on the PolitiFact dataset increased its accuracy on non-ambiguous verdicts from 75% to 89%. The accuracy on "true" claims jumped by +13.62 absolute points \cite{quelle2024perils} • The Fact-Check-Then-RAG method improved Llama 3 70B's accuracy on the PubMedQA dataset from 60.60% to 73.60% (+13.0 absolute points) by using fact-checking results to guide retrieval \cite{tran2024leaf} • An RAG pipeline using **Mixtral** achieved a **0.780 F1-score on the 'Refuted' class** on	3

	the Averitec development set \cite{singhal2024evidence}	
Hybrid and Multi-Agent Systems are More Effective	- Compared to strong LLM baselines like Flan-T5 and ChatGPT, the LoCal multi-agent system shows an average performance improvement of up to **7.75%** in the gold evidence setting and up to **6.17%** in the open book setting. \cite{ma2025local} - The hybrid SLM+LLM ARG network improved F1-score over its BERT-only baseline by +3.1 absolute points on the Weibo21 dataset (78.4% vs. 75.3%) \cite{hu2024bad} - The PACAR framework, with specialized agents, outperformed a general ChatGPT baseline by +16.9 absolute points on HOVER 4-hop claims (72.61% vs. 55.72%) \cite{zhao2024pacar} - The FACT-AUDIT adaptive multi-agent framework demonstrated superior evaluation robustness over static, single-agent pipelines by dynamically	4

	assigning roles \cite{lin2025fact}	
Automated Feedback Mechanisms Reduce Hallucinations	• **LLM-AUGMENTER** system, using a BM25 knowledge consolidator and automated feedback, improved the **KF1 score to 37.41 over the GPT KF1 score of 31.33** \cite{peng2023check} • The Self-Checker framework, an internal verification loop, improved label accuracy on the BINGCHECK dataset from 21.0% (ReAct baseline) to 63.4% \cite{DBLP:conf/naacl/LiPGGZ24/self-checker} • Medico's multi-source evidence fusion and correction loop improved hallucination detection F1-score by +34.4 points over its baseline on the **HaluEval** dataset \cite{zhao2024medico} • The Visual Fact Checker uses object detection and VQA models as automated "tools" to verify and correct initial caption proposals, significantly reducing hallucinations in detailed image captions	4

	\cite{ge2024visual}	
Fine-tuning on Synthetic Data Boosts Performance	• The MiniCheck-FT5 model, trained on generated synthetic data, achieved a Balanced Accuracy of 74.7% on the LLM-AGGREFACT benchmark, a +4.3 absolute point improvement over the previous state-of-the-art. An ablation study showed that removing this synthetic data caused the model's performance to drop by -14.8 absolute points, underscoring its critical importance \cite{tang2024minicheck} • FACT-GPT was trained on a synthetic dataset of contradicting, entailing, or neutral claims generated by GPT-4, which enabled a smaller, specialized LLM to match the claim-matching accuracy of larger models. \cite{choi2024fact}	2