http://www.casestudiesjournal.com/
Impact Factor: 4.428

Using AI Tools to conduct Investment Research on Emerging Technology

Author's Details:
[1] **Joshua ARJANTO** [2] **CHEN Xiaojing** [3] **Shiyou HU** [4] **Asher Ethan KOH**
[5]**Muhammad Khaizuran Bin MOHAMAD ROSLE** [6] **Lawrence LOH** [7] **Tim ZHANG**
[1] [2] [3] [4] [5] [6]National University of Singapore
[7] Edge Research Pte. Ltd.

1. Abstract

This case study examines how large language models (LLMs) including ChatGPT, Gemini, and DeepSeek can systematically augment the investment research process across five emerging-technology sectors: Artificial Intelligence, Robotics, Quantum Computing, Space, and Fusion. The analysis evaluates LLMs not as end-to-end automation, but as accelerators within a governed, human-in-the-loop workflow. Across sector onboarding, multi-lingual discovery, numeric extraction, and first-draft synthesis, LLMs reduced mechanical workload by broadening source coverage and enabling faster iteration. However, the evidence indicates that human judgment remains indispensable for causal reasoning, credibility assessment, evidence weighting, and final investment interpretation; LLMs shift analyst effort but do not replace it. To improve evidence hygiene at scale, the case describes a Python-based fact-checking pipeline that automates link reachability, publication-date extraction, and AI-generation risk screening, later extended into a low-code web interface. The workflow reduced verification time from 4.3 hours to 1.25 hours per 50 links (a 71% reduction), with projected savings of over 90% under full integration of headless browsing and GPTZero API scoring. In parallel, a structured source database with over 250 LLM-cited records (graded by recency and credibility) enabled reproducible evaluation of model outputs and surfaced common failure modes such as hallucinations, outdated citations, and contextual drift. Overall, the case study demonstrates that LLMs are most effective as structured force multipliers when paired with rigorous verification tooling and human oversight

Keywords: *investment research, large language models, due diligence, source verification, hallucination risk, reproducibility, artificial intelligence*

2. INTRODUCTION & PROBLEM CONTEXT

2.1 Case Study Scope (Emerging Tech: AI, Robotics, Quantum, Space, Fusion)

This case study evaluates how AI-based research tools can be used to gather information and support investment research across five emerging-technology sectors from G20 countries: Artificial Intelligence, Robotics, Quantum Computing, Space Technologies, and Nuclear Fusion. The objective is to assess whether recent advances in LLMs can improve efficiency, quality, and scalability of investment research while maintaining accuracy, credibility, and recency—qualities expected in professional investment analysis (Defend & Mortier, 2025).

Unlike a methodological study of LLMs themselves, the project primarily focused on using these AI tools to conduct real research on the aforementioned sectors and to evaluate the credibility of AI-assisted

findings in practice. Alongside the primary research effort, a secondary objective was to attempt the development of a sustainable fact-checking process; eventually leading to a prototype tool (Appendix D) capable of validating AI-generated investment research report outputs through link reachability, recency analysis, and basic source-credibility assessment using AI generation scores.

During the initial phase, three leading LLM-powered platforms were tested, which were Google Gemini, OpenAI ChatGPT, and DeepSeek. This familiarisation period was used to understand each LLM's capabilities, interfaces, and limitations across tasks such as document reading, web extraction, image parsing, citation generation, and long-context reasoning. Outputs were then used in a structured weekly cadence to construct sector value-chain analyses and business-impact briefings for each vertical, while simultaneously refining the verification workflow that underpinned later deliverables.

The second component of the case expands into the development of a prototype fact-checking tool to automate parts of the verification process. This tool addresses a recurring challenge in AI-assisted research: ensuring source reachability, recency, and credibility in automatically retrieved references. Together, these two streams (LLM-assisted sector research and verification tooling) provide the basis for actionable recommendations on how LLMs can be integrated into investment research workflows in a sustainable and responsible manner.

2.2 Baseline: The Traditional Investment-Research Workflow

In a traditional investment research framework, analysts typically engage in a manual process of collecting, verifying, and synthesising information from various primary and secondary sources to derive insights about sectors or value chains. This process often involves reviewing market research reports, analyst publications, and leveraging third-party data providers such as Bloomberg and S&P Capital IQ for company and market data. Analysts progress through a structured yet highly manual sequence of stages to build domain understanding, construct market views, and eventually develop actionable insights (Collin, 2025).

The workflow generally follows sequential stages. Analysts start with sector onboarding, building foundational literacy through primers and reference materials. They then conduct market sizing, gathering quantitative and qualitative data to assess the sector's attractiveness. Next, they map the value chain and screen companies to form a structured universe and shortlist. This is followed by company deep dives, analysing financials, strategy, and valuation to gauge traction and competitive position. The process concludes with report writing and review, where insights are synthesised into a final output and validated for accuracy and rigour (Twin, 2025).

Across all stages, recurring friction points include long hours of manual research, fragmented data, conflicting estimates, time-intensive extraction from PDFs and transcripts, and the compilation and reconciliation of materials. Analysts also spend significant effort tracking the recency and credibility of each source. Completing a full five-stage process for a single sector typically requires two to three months of full-time analyst work to reach actionable recommendations. These pain points define the baseline process as slow, costly, and repetitive, providing a clear benchmark for comparison against an LLM-assisted workflow.

2.3 Guiding Questions (Augment vs Replace)

This case study explores how LLMs and related AI tools can enhance the traditional investment research workflow, focusing on the balance between augmentation and replacement. While manual processes offer rigour and traceability, they are often slow, labour-intensive, and limited by human bandwidth, particularly in parsing lengthy documents, reconciling conflicting data points, and maintaining consistency across sources. The central question is whether AI can accelerate these workflows without compromising analytical depth or evidence quality (LexisNexis, 2024).

This study considers four guiding questions:

1. Where in the investment research process (such as sector onboarding, market sizing, value-chain mapping, and company deep dives) can AI tools most effectively augment analyst work?

2. How do AI-assisted workflows compare with traditional manual research in terms of speed, coverage, and evidence quality?

3. What limitations or risks arise when integrating AI outputs, particularly around hallucinations, citation accuracy, or contextual loss?

4. What is the right balance between automation and human oversight to ensure credible, reproducible insights?

These questions form the analytical foundation for later comparisons between traditional and LLM-assisted approaches and inform process recommendations on where AI can credibly augment (and where it should not replace) core analyst work.

2.4 Success Criteria (Quality, Recency, Cost/Time, Reproducibility, Hallucination Risk)

To evaluate whether AI tools can credibly augment the investment research process, this case study defines a clear set of success criteria across five dimensions. These criteria measure both the efficiency gains and the quality outcomes of LLM-assisted workflows compared to traditional manual research (Niel et al., 2024).

Quality and Accuracy:

Outputs must be factually correct, internally consistent, and supported by credible primary or secondary sources. A successful AI-assisted workflow should produce findings that are at least as accurate as traditional manual research, while making it clear where each fact or number came from. Every statement should be traceable to an identifiable source, and the reasoning behind key conclusions should be easy for another analyst to review and verify (Shah, 2025).

Recency and Coverage:

Given that emerging technology sectors evolve rapidly, the ability to retrieve and synthesise up-to-date information is critical. Success is measured by whether AI tools can surface recent data points (e.g., latest financials, funding rounds, product launches, regulatory actions) and provide comprehensive coverage across sub-sectors and geographies.

Cost and Time Efficiency:

AI augmentation should reduce the time and labour required to complete core research tasks such as data extraction, summarisation, and drafting. Improvement is quantified by comparing average task duration and analyst hours between manual and AI-assisted workflows, without sacrificing quality or rigour.

Reproducibility and Consistency:

A reliable workflow should produce consistent outputs when repeated under similar prompts or datasets. Success is defined by the ability to replicate findings, maintain standardised formatting, and avoid subjective drift across different analysts or model sessions.

Hallucination and Risk Management:

Reducing factual hallucinations and citation errors is essential for credibility. The benchmark for success is a verifiable output where every quantitative figure and factual claim can be traced to a live, reputable source (Moore, 2025). Hallucination mitigation is evaluated through cross-checks, link logging, and manual verification steps.

Together, these criteria form the evaluative framework for subsequent sections. They underpin comparisons between traditional and LLM-assisted research and inform the process recommendations that operationalise an analyst-facing, human-in-the-loop model at scale.

3. METHODOLOGY OVERVIEW

To maintain comparability of outputs, a common value-chain framework was applied across all sectors. The value-chain schema was segmented into “Suppliers” (inputs and tooling), “Upstream” (core hardware/compute), “Midstream” (platforms and control layers), and “Downstream” (end applications). Parameters, definitions, and standards of evidence were held constant across the entire case study period.

3.1 Tools Used (ChatGPT, Gemini, DeepSeek; Python Notebooks)

Tooling was kept lightweight and task-specific. Generally, Gemini was used for first-pass breadth and recency (news, announcements, funding, policy); ChatGPT was used for document reading and structured extraction from filings/transcripts and for producing clean comparison tables; DeepSeek was used for Asia/China enrichment and multilingual lookups. Deeper capability comparisons are deferred to the model-comparison section.

Python notebooks were used to develop and validate the verification prototype, enabling rapid iteration, clear step-by-step structure, and tight integration with Python's library ecosystem. The prototype's objectives, architecture, and results are detailed in the verification section.

3.2 Data Governance: Citations, Link Logging, Source Recency Checks

All sources found and collated were continuously recorded in a source-database spreadsheet capturing features such as URL, date accessed, value-chain position, and notes. Each source was assigned a credibility score and a recency rating to prioritise up-to-date information. Cross-source triangulation was required for key facts, and unresolved conflicts were explicitly annotated. This governance ensured traceability, reproducibility, and clear separation between model-suggested leads and human-verified evidence.

For a full view of the database (fields, scoring rubrics, examples, and how records are used in analysis), see Appendix A: Source Database.

4. LLM CAPABILITIES & CROSS-MODEL COMPARISON

4.1 Model Snapshots & Versions Used (With Dates)

Over the case period, three systems were used consistently, each playing a distinct and complementary role.

ChatGPT (GPT-5 Thinking family)

Served as the main analytical environment, used daily for drafting, document ingestion, numeric extraction, reconciliation, table construction, and code generation. Stability and the ability to handle long, structured documents made it the backbone of the workflow.

Gemini (Deep Research; Flash 2.5)

Employed during web-grounded research sprints, especially when broad, up-to-date sweeps of sources were needed. Recency performance made it well-suited for gathering fast-moving market information, regulatory changes, and technology updates. Its behaviour in recency-focused retrieval is consistent with descriptions in the Gemini technical report (Google DeepMind, 2024).

DeepSeek (latest public release during project window)

Used selectively when Chinese-language materials were central: corporate disclosures, industry standards, policy documents, and Chinese market news. Outputs were then normalised, translated more precisely, and stylistically harmonised.

Across the case period, this division of labour stayed relatively stable: ChatGPT handled most analytical and editorial work; Gemini provided breadth and timeliness for evidence gathering; DeepSeek supported China-specific discovery where English analogues were limited.

4.2 Comparative Strengths

In practice, ChatGPT proved most dependable for "closed-book" document work: reading PDFs, performing unit-faithful numeric extraction, reconciling conflicting figures across sources, and turning raw notes into clean tables, figures, and code. Gemini excelled when breadth and recency mattered; when seeded with focused prompts, it returned balanced, source-linked overviews and surfaced counter-arguments that improved report neutrality. DeepSeek's advantage was recall and nuance for China-specific materials, mapping

Chinese technical terminology to defensible English equivalents before final editorial passes in ChatGPT (Refer to *Table 1*, Appendix E: Tables).

4.3 Cost, Latency, Reliability (Hallucination/Coverage Issues)

Task turnaround was controlled by specialising tasks across models. ChatGPT handled iterative drafting and document-grounded analysis; it was responsive for short edits but slower on long, structured outputs such as large tables and multi-document synthesis. Gemini was used in scheduled, web-grounded sweeps for new topics or updates; latency was batch-oriented and longer, but returns were well-organised and citation-rich digests. DeepSeek was used on demand for Chinese-language retrieval, delivering moderate-fast responses on targeted lookups.

Reliability varied by task. ChatGPT showed the lowest hallucination rate in document-grounded tasks, yet it could over-generalise on broad topics without fresh sources. Gemini's breadth occasionally introduced over-aggregation, near-duplicate links, inconsistent "as-of" dates, link rot, and unfiltered vendor material that could inflate claims unless prompts enforced strict dating and provenance. DeepSeek sometimes carried translation/scale drift when converting Chinese language to English language figures (e.g., mapping "亿" imprecisely), and occasional A/H-ticker mismatches, requiring normalisation before integration.

Mitigations included source-backed prompting, arithmetic/identity checks for quantitative claims, top-N caps and de-duplication for Gemini results, explicit date/provenance constraints, and a glossary/style-normalisation pass before integration. Similar reliability patterns have been documented in surveys of LLM hallucination and multilingual alignment (Ji et al., 2023; Zhang et al., 2024).

4.4 Failure Modes

Across models, distinct error patterns required disciplined controls. ChatGPT/GPT-5, which was used for analysis and table construction, occasionally produced plausible-looking numeric hallucinations when parsing filings or scraped tables, such as pairing a quarter-end enterprise value with TTM denominators, reading parentheses in disclosures as negative EBITDA rather than stylistic formatting, and drifting between units or currencies when OCR artefacts were present. Gemini (Deep Research/Flash 2.5) was strongest on recency, but its breadth sometimes came with source-hygiene issues: over-aggregation of near-duplicate links, inconsistent as-of dates across citations, link rot, and unfiltered vendor material that could inflate claims unless prompts enforced strict dating and provenance. DeepSeek, used for Chinese-language discovery, showed high recall yet introduced terminology and scale drift during translation (e.g., mapping "亿" imprecisely to "billion"), occasional A/H-ticker mismatches, and RMB-to-USD conversions without explicit FX references, all of which needed normalisation before integration. Similar Chinese language to English language scaling inconsistencies are highlighted in multilingual LLM alignment research (Zhang et al., 2024).

4.5 Proposed Solution: Workflow Design & Control Framework

As the case progressed, it became clear that no single model could handle every part of the workflow reliably. Instead, a process emerged that aligned tasks to model strengths while managing weaknesses. Work fell into three phases: first, collecting external sources with a wide net; second, extracting document-grounded facts; and third, reconciling everything into a consistent dataset. Gemini was used to collect breadth and recency, and DeepSeek was used to retrieve specific Chinese-language materials. ChatGPT handled reconciliation, table construction, and final synthesis.

This structure became the most stable way to keep analysis accurate as the dataset grew. When ChatGPT produced numbers that appeared reasonable, they were still checked against the filing by logging units in the prompt and running deterministic arithmetic checks before accepting derived metrics. When Gemini returned overlapping links or mixed evidence from different reporting dates, results were controlled by limiting domains and forcing explicit "as-of" dates for each citation. DeepSeek's translations were checked for unit fidelity, ticker identity, and RMB-to-USD conversions tied to explicit FX references.

Overall, a general workflow paired with targeted controls proved most reliable. It allowed analytical synthesis to remain in ChatGPT, recency sweeps to remain in Gemini, and Chinese-language visibility to remain in DeepSeek, while reducing failure modes before final deliverables.

4.6 How Usage Evolved as Newer Models Released

Tool usage shifted as both the case and available tools evolved. At the start, there was heavier reliance on GPT-4, which was capable of strong reasoning and drafting but struggled to surface the most recent web materials. Because early-stage research required wide and timely evidence gathering, Gemini (Deep Research/Flash 2.5) became the tool for recency-based sweeps and on DeepSeek for Chinese-language sourcing.

This changed when GPT-5 was released. Its improvements in numeric extraction, table construction, document reconciliation, and code-assisted analysis directly matched the needs of the second half of the workflow, which had shifted from conceptual mapping to quantitative estimation and validation. Once GPT-5 proved more consistent across these tasks, it became the central model for synthesis and final deliverables. Gemini remained valuable for periodic web updates, and DeepSeek continued to provide Chinese market visibility, but the bulk of the analytical effort migrated to GPT-5. This transition reduced rework and accelerated the turnaround time for tables, figures, and appendices.

5. TRADITIONAL VS LLM-ASSISTED RESEARCH

This section compares the established, manual investment-research workflow used across industry with the emerging LLM-assisted workflow on trial. Drawing on sectoral research experiences across AI, Robotics, Quantum, Space, and Fusion, the analysis evaluates how LLMs reshape effort profiles, output quality, and analyst time allocation. While the benefits of LLMs are meaningful, particularly in accelerating early-stage scoping, data processing, and first-draft generation, these gains are counterbalanced by persistent limitations in reasoning, judgment, and credibility verification. The comparison below clarifies where LLMs augment the research process versus where human capabilities remain central and irreplaceable.

5.1 Effort & Time Estimates for Core Tasks (Manual Baseline vs LLM-Assisted)

5.1.1 The Traditional Investment Research Workflow

Industry practice for investment analysts typically relies on a structured six-phase workflow that spans 8-12 weeks for a comprehensive sector or thematic report. Each phase is sequential and labour-intensive, requiring multi-disciplinary expertise across data sourcing, financial modelling, narrative synthesis, report drafting, and client communication.

1. **Scoping & Design (2-4 days)**

   Analysts begin by defining the research question, establishing scope, and framing hypotheses. This stage is inherently iterative and often shaped by initial conversations with stakeholders, past coverage, and internal knowledge. Manual workflows require analysts to independently explore the opportunity set and identify gaps.

2. **Data Collection (1.5-2 weeks)**

   This phase is the heaviest in terms of raw hours. Analysts manually gather macro statistics, regulatory updates, company filings, earnings transcripts, market-share data, and expert insights, often dispersed across paywalled sources. Data cleaning and organisation are major sub-tasks (Vipond, 2025). Human judgement is required to determine relevance, reconcile conflicting data, and ensure source quality.

3. **Analysis & Modelling (2-3 weeks)**

   Traditional workflows involve building financial models, benchmarking peers, constructing sensitivity scenarios, and validating assumptions. This requires technical modelling expertise and domain understanding. Errors can propagate easily, so analysts spend long hours verifying calculations and cross-checking inputs.

4. **Synthesis & Narrative Development (1.5-2 weeks)**

   Analysts translate data and models into coherent investment theses, narratives, and visual exhibits. Writing is performed from scratch and often requires significant redrafting (CFA Institute, 2020). This stage demands an ability to spot patterns, weigh materiality, and connect disparate insights into a persuasive argument.

5. **Review & Publication (3-5 days)**

   Compliance, peer review, fact-checking, formatting, and editorial refinement occur here. The process is slow due to manual cross-referencing and institutional sign-off requirements.

6. **Dissemination & Follow-Up (2-3 days)**

   Finally, analysts present findings to clients and investors, handle Q&A, and update coverage based on market feedback. Follow-up is periodic, reflecting manual monitoring of news, filings, and sector developments (Grata, 2025).

This traditional workflow is thorough and high-confidence but slow, resource-intensive, and constrained by human bandwidth, especially during data-heavy phases.

5.1.2 The LLM-Assisted Investment Research Workflow

Experimentation with LLM tools produced a modified workflow with a materially reduced timeline of 3-6 weeks. The reduction is achieved not by eliminating steps but by accelerating specific tasks within each phase. The workflow retains the six phases but redistributes effort.

1. **Scoping & Design (1-3 days)**

   LLMs help analysts shape hypotheses, surface initial angles, and identify data gaps quickly. By prompting models with preliminary sector context, analysts can obtain landscape maps, potential sub-themes, and early framing suggestions. This reduces the time spent on exploratory scoping by roughly half.

2. **Automated Data Collection (3-6 days)**

   LLMs can cut the time required for data harvesting by enabling rapid summarisation of company filings, web sources, transcripts, and regulatory documents. AI-assisted extraction and classification can reduce tagging and basic organisation work that traditionally consumes analyst hours. This is where the largest efficiency gains were observed (60-70% time reduction).

3. **Assisted Analysis & Modelling (1-1.5 weeks)**

   LLMs generate preliminary ratio analyses, scenario outlines, peer benchmarks, and sensitivity structures, which analysts then validate and refine. LLMs do not replace modelling expertise but can

accelerate the production of workable starting points (Pop et al., 2024). Analysts still review assumptions, ensure consistency, and correct logic. Time savings here averaged 40-50%.

4. **AI-Drafted Synthesis (4-6 days)**

   LLMs produce first-draft narratives, structure key messages, and generate draft tables or exhibit text. Analysts then add domain judgement and sector-specific nuance, tighten claims, and correct inaccuracies (Madanchian & Taherdoost, 2025). This stage typically delivers strong efficiency gains (50-60%).

5. **Review & Compliance (3-5 days)**

   Human review remains essential. LLMs can assist with tone harmonisation, citation formatting, and consistency checks, including flagging contradictory statements or unclear reasoning. Time savings are more modest here (around 30%), partly due to compliance requirements that cannot be fully automated.

6. **Continuous Update Loop (Real-time)**

   LLM support can also enable faster monitoring of filings, macro releases, and news. Instead of periodic batch updates, analysts can work with lower-latency prompts and alerts, improving responsiveness and enabling ongoing refinement of outputs.

Across phases, LLMs reduce time spent on repetitive processing and allow more attention to be directed toward higher-order reasoning rather than mechanical tasks.

5.1.3 Comparative Time Impact

The combined effect of LLM assistance is a 40-60% reduction in overall turnaround time (Refer to *Table 2*, Appendix E). Tasks that previously required highly manual effort, including data gathering, summarisation, preliminary modelling, and first-draft writing, experience the greatest acceleration. However, human-dependent portions of analysis, such as judgment, credibility checks, and final decision framing, see smaller reductions because they cannot be reliably delegated to LLMs.

5.2 Output Quality & Coverage: What Improved, What Still Needed Human Judgment

5.2.1 Improvements Observed with LLM Assistance

1. Expanded coverage and breadth

LLMs enabled analysts to review more companies, geographies, and thematic sub-topics than would be feasible manually (Cognizant, 2025). In sectors with many niche players, LLM-based summarisation can support faster screening, improving comprehensiveness and reducing blind spots.

2. Faster synthesis and clearer structure

Across sectors, LLMs can produce well-organised draft narratives and transform notes, tables, and models into coherent prose. This is particularly useful in domains with complex value chains and dense technical context, where organising information is a major bottleneck.

3. Pattern recognition and anomaly surfacing

LLMs can help surface KPI shifts, peer-trend divergences, and recurring themes across filings or earnings calls. This can shorten the time required to identify cross-company signals that would otherwise require extensive manual reading.

4. Real-time, always-on monitoring

LLM-assisted alerts can strengthen the ability to track regulatory updates, new research releases, macro policy shifts, and quarterly announcements, reducing latency in follow-up analysis (AWS, 2024).

5. Improvements in communication quality

LLMs can improve clarity, flow, and tone consistency across written outputs, accelerating the production of presentation-ready documents.

5.2.2 Areas Where Human Judgment Remained Essential

1. Judgment about materiality

LLMs can list many factors, but they often struggle to determine which drivers are truly valuation-relevant. Analysts must weigh magnitude, likelihood, and strategic relevance (Kang & Liu, 2023).

2. Causal reasoning and inference

LLMs frequently describe correlations but cannot reliably infer causality. Explaining performance changes, identifying operational bottlenecks, and translating technical change into business-model implications remain judgement-intensive tasks.

3. Credibility checks and contextual grounding

Decision-grade research relies on triangulation across primary sources, expert input, and contextual knowledge. LLM outputs require verification and cannot substitute for credibility assessment.

4. Handling ambiguity and grey areas

Frontier sectors often involve sparse or contradictory data. LLMs may smooth over gaps with generalisations. Analysts must adjudicate uncertainty and frame risks and opportunities appropriately.

5. Accountability and ethical standards

Analysts are accountable for their calls, valuations, and recommendations. LLMs do not hold responsibility and cannot ensure compliance adequacy. Regulatory standards, particularly around forward-looking statements and risk disclosure, necessitate human oversight (Constantinescu & Kaptein, 2025).

Overall, LLMs accelerate the "input" side of research (data, structure, speed) but humans remain indispensable on the "output" side (judgment, credibility, and accountability).

5.3 Takeaways: Where LLMs Augment, Where They Cannot Replace

5.3.1 Where LLMs Augment Analysts Effectively

LLMs enhance workflows in speed and scale, large-corpus pattern surfacing, and first-draft structuring. These strengths reduce time spent on repetitive tasks and support faster iteration.

5.3.2 Where LLMs Cannot Replace Analysts

LLMs remain insufficient for materiality judgements, causal explanation, credibility assessment, working under ambiguity, and accountability for final outputs. For these reasons, LLMs function as accelerators rather than substitutes for analyst expertise.

6. SOURCE DATABASE ANALYSIS

This section documents the consolidated evidence log used throughout the project and summarises key patterns across credibility, recency, and AI-generation scores. Between the four-moth case study period, sources were logged approximately every two weeks, resulting in more than 250 records across five sector verticals (AI, Robotics, Quantum, Space, Fusion). After data cleaning and standardisation, each vertical contributes 50 data points of "ready" samples (valid links with complete fields) for comparative analysis by LLM and model.

6.1 Database Contents and Analysis Preparation

Each record stores the minimum information required to retrace a claim. This includes the source publisher and type, a brief one-liner content summary, credibility and recency scores (1-5), GPTZero AI/Mixed/Human likelihood scores (used only as a triage signal), and the LLM/model that surfaced the link.

This structure supports analysis of three signals: quality (credibility/recency), AI-screening risk (GPTZero), and link hygiene (reachability).

Labels were normalised (e.g., folding "Gemini DeepResearch" into Gemini; unifying GPT-5/4o variants), dates were coerced, and fields were validated. For valid links, the AI-score triplet must be present and must sum to 100%. For invalid links, AI scores remain blank. "Ready" rows require a reachable link plus complete, internally consistent credibility, recency, date, AI scores, and canonical LLM/model labels. "Non-ready" rows remain in the raw workbook for traceability.

For the full field dictionary, scoring rubrics (credibility/recency), normalisation rules, and QA checks, see Appendices. Section 6 focuses on presenting the key findings.

6.2 Credibility and Recency Signals

This subsection evaluates the quality of sources surfaced by the LLMs using two rubric-based scores: Credibility (1-5) and Recency (1-5). Rubrics were defined in advance and applied consistently across all sectors. Average credibility and recency were compared across vendors and model variants on the 250+ "ready" rows (50 per sector). The figures titled "*Average Credibility by LLM/Model*" and "*Average Recency by LLM/Model*" summarise the results (Appendix E). The two scoring frameworks are reproduced in *Table 3 (Recency rubric)* and *Table 4 (Credibility rubric)* (Appendix E).

6.2.1 Credibility Analysis

On credibility, ChatGPT leads with an average score of 4.15, followed by DeepSeek (3.90) and Gemini (3.62). At the model level, GPT-5 posts the highest average (4.19), while DeepSeek V3 and GPT-4o cluster around 3.90, and Gemini Flash 2.5 trails at 3.62. This pattern is consistent with workflow observations: ChatGPT more often surfaced primary filings and high-reliability outlets, whereas Gemini's breadth occasionally included trade press and company blogs that were scored as credible but not authoritative (Reynolds, 2025). DeepSeek's credibility average benefited from frequent pulls of official Chinese sources but was moderated by mixed-quality secondary reporting (Appendix F: Figures 1 and 2).

6.2.2 Recency Analysis

For recency, Gemini edges out with an average of 3.74, ChatGPT follows at 3.64, and DeepSeek lags at 3.10. Model-level results mirror this: GPT-5 and Gemini Flash 2.5 both average 3.74, while DeepSeek V3 and GPT-4o sit around 3.10. In practice, Gemini tended to retrieve more recently stamped news posts and updated documentation (Jayaraman, 2025); DeepSeek's lower recency reflects a higher share of evergreen technical pages and older program announcements from Chinese portals. (Appendix F: *Figures 3 and 4).*

6.2.3 What This Means for Routing and Prompts

- **When freshness matters** (*policy changes, company news, product updates*), Gemini can be prioritised, or GPT-5 can be used with explicit recency constraints (e.g., "past 90 days") alongside programmatic date extraction and checks.
- **When source authority matters** (numbers used *in slides and investment theses*), ChatGPT/GPT-5 can be prioritised to raise the hit-rate of primary filings, regulator releases, and top-tier publications. DeepSeek can be retained for Asia coverage and recall, with validation against primary documents.
- **For China/Asia recall specifically**, DeepSeek V3 can be used to surface entities and local documents, then pair with a recency filter and cross-verification.

6.2.4 Caveats

These results reflect averages rather than hypothesis tests. Scores follow the stated rubrics and the sector mix over the case study period; distributional differences (e.g., medians, interquartile ranges) are not shown here and may narrow apparent gaps. Some observed differences may reflect usage effects rather than model effects, including prompt framing, source filters (e.g., "use primary sources only"), and task allocation across tools. These findings are therefore treated as routing heuristics rather than absolutes: tool choice should be matched to task, automated date extraction should be enforced, and final source selection should remain judgement-led.

6.3 AI-Generated Content Risk Score Signals

This subsection interprets GPTZero scores logged for each source as a **triage signal** rather than a truth test. GPTZero classifies text into three proportions (AI, Mixed, Human) based on detector features. These outputs are used only to flag items for closer review; final judgment remains human.

The figures titled "*AI-Score Composition by LLM*" and "*AI-Score Composition by Model*" show the average percentage split of AI / Mixed / Human likelihoods across the 50 "ready" sources per sector (valid link with complete metadata). Bars are grouped by (i) the assistant used (ChatGPT, DeepSeek, Gemini) and (ii) model families (GPT-5, GPT-4o, Gemini Flash 2.5, DeepSeek V3). (Appendix F: *Figures 5 and 6*)

6.3.1 AI Score By LLM

The distributions differ meaningfully across assistants (*Figure 5,* Appendix F)

- **ChatGPT**: Highest Human share at approximately 78%, with roughly 10% AI and 11-12% Mixed. In practice, ChatGPT-surfaced links more often pointed to primary documents or detailed technical posts, consistent with the higher Human band shown in the figure.
- **DeepSeek**: Mid-range profile with approximately 70% Human, 16–17% AI, and 12-13% Mixed. These links included a mix of vendor notes, technical blogs, and trade press; tehy were generally suitable for screening but were more frequently flagged for second-pass review than ChatGPT.
- **Gemini**: Lowest Human proportion at approximately 56%, with the highest AI (around 26%) and Mixed (around 18%) bands. Many of these were polished secondary sources or syndicated pages, which are treated as "read but verify" items.

In routine review, items with larger AI/Mixed shares coincide with promotional copy, aggregated explainers, or reposts; accordingly, these items receive more manual spot checks (Patel, 2025).

6.3.2 AI Score By Model

A similar pattern appears at the model level (*Figure 6,* Appendix F*).*

- **GPT-5**: Cleanest profile with approximately 83% Human, around 9% AI, and around 8% Mixed; the smallest combined AI+Mixed band in the cohort.
- **GPT-4o**: Higher Mixed (around 32%) and around 16% AI, with Human around around 53%. This aligns with outputs that read polished and templated, which GPTZero often marks as Mixed.
- **DeepSeek V3**: Human around 71%, AI around 16%, Mixed around 13%; close to ChatGPT overall, but with slightly more AI/Mixed than GPT-5.
- **Gemini Flash 2.5**: Human around 56%, AI around 26%, Mixed around 18%, consistent with the assistant-level pattern.

Where GPT-5 is the retrieval model, fewer rows trigger a "needs extra scrutiny" flag from the detector. GPT-4o and Gemini Flash 2.5 yield higher Mixed/AI-looking prose on average, so more second-pass reading is budgeted for those batches.

6.3.3 Operational Use Of AI Screening Signals

GPTZero is used strictly as a triage aid. When a source returns an AI score of approximately 25% or higher, or a Mixed score of approximately 30% or higher (particularlly when combined with a Credibility rating of 3 or below), the row is routed to a second-pass human review. Primary sources such as regulatory filings, regulator websites, and OEM datasheets remain in scope regardless; if templated language triggers higher scores, the reason is annotated.

GPTZero is not treated as a gatekeeper. Polished human prose may be flagged as "Mixed," and lightly edited AI text can sometimes register as Human. No source is added or removed on the detector's output alone.

In practice, combined with credibility and recency ratings, the detector helps redirect attention to higher-risk items. Batches produced with GPT-5/ChatGPT generally require fewer follow-ups, while Gemini Flash 2.5 and GPT-4o batches warrant additional spot checks where the AI/Mixed signal trends higher.

7. BASELINE MANUAL PROCESS

This section documents the as-is fact-checking workflow that was used before automation. Its purpose is to ensure that LLM-produced claims and citations are traceable, recent, and credible, and to establish a measurable baseline for the prototype.

7.1 Procedure: GPTZero Checks, Manual Link Clicking, Recency Validation, Spreadsheet Log

LLM "deep research" outputs typically contain approximately 50 hyperlinks per report, plus additional references surfaced via ad-hoc queries. Each cited source is validated through a consistent, link-by-link process.

1. ***Reachability check (manual clicking)***: Open the URL to confirm it resolves (no 404/403/timeout) and is not gated by an impenetrable paywall.
2. ***Content validation***: Read the page and verify that the quoted numbers/claims in the LLM output actually appear in the source and are interpreted correctly (units, currency, period, context). Any mismatch or hallucinated citation is flagged and recorded in the slide-deck notes.
3. ***Recency & credibility***: Identify an on-page date (publication or last updated) and the publisher type (e.g., regulatory filing, company press release, reputable trade or analyst outlet, blog, forum). Primary sources are preferred; low-credibility items are downgraded or excluded.
4. ***AI-generation screening (GPTZero):*** Copy page text into GPTZero to obtain an AI-generation score (percentage likelihood of AI-generated content). This is treated as a risk signal (not an absolute filter) to trigger deeper review where scores are high or inconsistent with source type.
5. ***Structured logging***: Record all outcomes in the fact-checking spreadsheet (source database) with fields including URL, title, publisher, brief content summary, access date/time, HTTP status, detected page date, GPTZero score, LLM/model used, and any extra notes.

7.2 Time Profile & Pain Points

The manual process is time-intensive, averaging approximately 5-8 minutes per link from click to log entry. With around 50 links per report, a single deep-dive pass can consume roughly 5-6 hours of analyst time before any rechecks. In practice, a meaningful share of effort is lost to content mismatch (cases where the numbers cited in the LLM output do not appear on the page or differ in vintage, units, or scope) as well as duplication when multiple URLs resolve to the same canonical article after redirects. Logging is another source of friction: manual copy-paste introduces inconsistencies that slow later audits and revalidation. Together, these

issues make the baseline slow, repetitive, and error-prone, and they limit how frequently recency checks can be refreshed. This workflow is therefore a clear target for optimisation to reduce analyst time and streamline research operations.

7.3 Justification for Automation and Prototype Design

Given these pain points, automation is justified to address three checks required for fact-checking at scale: link validity, date of publication, and AI-generation risk. The prototype is designed as a staged Python workflow:

(i) Link-health pass: Request each URL, follow redirects to a canonical address, record status codes, and drop dead or duplicate links. This removes time spent on 404/403 loops and repeated articles.
(ii) Recency pass: Fetch the page and attempt to extract a publication or updated date from visible page text and common metadata (e.g., article:published_time, og:updated_time, date, last-modified). Where no reliable date is found, the script records "unknown" and flags the link for manual review.
(iii) Content screening pass: Extract readable text and call the GPTZero API (Appendix B) to compute an AI-generation score as a risk signal for deeper human review. Each step emits a structured record (URL, reachability, detected date, GPTZero score, and related fields) into the fact-checking database, creating a consistent audit trail.

By limiting scope to these three automated checks, the prototype reduces per-link cycle time, cuts duplication, standardises recency handling, and streamlines the LLM-assisted investment-research workflow. This reallocates analyst time toward judgement-based tasks (credibility and interpretation) rather than mechanical verification.

7.3.1 Why GPTZero?

To prioritise human review at scale, an independent signal is required to indicate whether a page's prose is likely AI-generated. GPTZero was adopted as that signal because it is trained on corpora containing both human and model-generated text and provides document- and sentence-level classifications. At a high level, GPTZero analyses writing statistics (e.g., perplexity and burstiness) and applies supervised classifiers to estimate whether text is Human, AI, or Mixed (Chen & Barlow, 2025). The returned score is logged alongside each URL's metadata in the fact-checking database.

Using "AI to detect AI" is valid in this workflow for a specific reason: modern generators leave statistical fingerprints that differ from the more irregular patterns of human writing. Detection models are trained explicitly on both classes to learn these contrasts, and vendors update them as generators evolve (Rosen, 2024). That said, GPTZero was strictly treated as a risk indicator, not a verdict. A high score flags a page for closer reading; it does not automatically exclude a source.

The approach also has limitations. False positives and false negatives are possible (e.g., heavily edited AI text may appear human; formal human prose may resemble AI) (Dhar, 2025). For this reason, GPTZero operates within a human-in-the-loop policy: credibility remains determined by source type, provenance, and direct claim verification. Operationally, this improves triage efficiency without outsourcing judgment, keeps decisions auditable, and preserves the integrity of the research record (Rosen, 2024).

8. PROCESS EFFICIENCY EVALUATION

In this section, the extent of how the revised workflow reduced analyst time at two stages of the workflow is summarised as such: (i) sector analysis supported by LLMs and (ii) fact checking using the full stack verification tool. Survey inputs, detailed assumptions, and calculation steps are documented in Appendix D.

8.1 LLM Assisted Sector Analysis

A conventional sector analysis completed by a single full-time analyst was used as a benchmark, using industry survey data on typical weekly hours and a 2.5 month project duration, and compared this with timesheet data collated during the study while working with LLM support over a 16 week period.

Under the manual benchmark, a full sector analysis requires approximately 445.63 analyst hours (Boyakhchyan, 2025; Toulon, 2025). In the LLM assisted setting, the findings showed that the effective single analyst load falls to 195.2 hours. Thus, this shows that the workflow saves 250.43 hours, a 56.2% reduction in analyst time for comparable decision useful outputs (value chain analysis, weekly briefs, company shortlists, and preliminary recommendations).

8.2 Fact Checking Workflow Using the Full Stack Tool

For verification, effort was normalised to cycle time per 50 links, based on lower and upper bound estimates for fully manual checking, and the same workload was then repeated through the tool.

The manual baseline averaged 4.3 hours per 50 links. After running the same task set through the verification tool (Appendix D) it reduced this to 1.25 hours per 50 links, a 71% reduction and a 3.44 times throughput gain. In a fully implemented configuration, cycle time is estimated to fall further to 0.375 hours per 50 links, which corresponds to a 91% reduction and an 11.5 times throughput gain, converting what was previously roughly a half day of checking into a workflow that can approach 20 to 25 minutes per 50 links.

9. PROCESS RECOMMENDATION (AUGMENT VS REPLACE)

This section translates the empirical findings into a practical operating model for LLM-assisted research. Instead of treating AI systems as end-to-end automation, a structured, human-in-the-loop workflow is proposed in which LLMs are assigned to clearly defined task types, and verification follows an outlined standard operating procedure (SOP).

Recommendations are framed into the two parts below:

1) Decision guide for when and how to use each model
2) Verification checklist for evidence hygiene

9.1 Decision Guide: When to Use Which Model

Experience from the case study indicates that the most robust outcomes arise from a task-to-model approach rather than attempting to use a single LLM for everything. At a high level, the following distinctions are made between:

- Closed-book, document-grounded tasks (numeric extraction, table building, reconciliation);
- Open-web, recency-sensitive tasks (new, fundings rounds, regulatory changes;
- Regional or multilingual discovery tasks (Asia/China company coverage); and
- Mechanical drafting and editing tasks (first drafts, re-phrasing, slide copy)

Within this framework, the recommended allocation is as follows.

a) Sector onboarding and landscape scans

For early-stage work, within this framework, the recommended allocation is as follows:

- **Primary**: Gemini Deep Research, prompted to surface diverse, recent sources with explicit citations.
- **Secondary**: ChatGPT, used to clean and normalise Gemini outputs into concise primers, glossaries, and value-chain narratives.

b) Market sizing, value-chain mapping, and company universe construction

For structured mapping tasks (CAGR ranges, value-chain layers, longlists of companies):

- **Primary:** ChatGPT for turning raw notes and filings into structured tables, value-chain diagrams, and company universes.
- **Secondary:** Gemini for topping up recency (e.g. latest funding rounds, new product lines) once a draft table is in place.
- **DeepSeek:** Used selectively to discover Asia-focused peers, Chinese tickers, and local exchanges, with all outputs subsequently normalised in ChatGPT to a standardised form.

Here, LLMs act as scaffolding tools: they accelerate the assembly of longlists and draft maps, but inclusion/exclusion decisions and value-pool judgements remain with the analyst.

c) Company deep dives and numeric extraction

Once companies are shortlisted, the workflow shifts to document-grounded analysis:

- **Primary:** ChatGPT for reading filings, investor presentations, and credible news; extracting revenue, capex, margins, multiples, and segment splits; and building comparison tables.
- **Gemini only** for filling obvious recency gaps (e.g. latest quarter not yet in internal materials), with any new numbers treated as *unverified* until checked against primary sources.

In this phase, ChatGPT is treated as a "smart spreadsheet assistant" rather than an oracle: all key fields (numbers, dates, units, currencies) are subject to the verification SOP in Section 9.2 before inclusion in final outputs.

d) Fact-checking, source logging, and risk triage

To ensure citation hygiene and effective link management, the following protocols are recommended:

- **The Python + Web Prototype** (Lovable) as the first pass for link reachability, recency extraction, and AI-generation risk signals;
- **Human analysts** to interpret edge cases flagged by the tool (e.g. missing dates, inconsistent numbers, high GPTZero AI scores)

Here, LLMs support the explanation layer (e.g. summarising a source, re-stating a claim in plain language). But the determination of whether a source is credible or fit for purpose remains a human decision.

e) Drafting reports, slides decks, and executive summaries

For narrative production:

- **Primary:** ChatGPT for drafting and iterating on executive summaries, section write-ups, and slide copy, seeded with verified tables and bullet points
- **Secondary:** Gemini only when more context is needed for framing (e.g. macro context, sectoral comparisons), and its outputs are cited and fact-checked as with any other source.

In summary, across all phases, the guiding principle is: Use LLMs to compress mechanical work, while reserving interpretation for human analysts.

9.2 Verification SOP Checklist

Given the centrality of evidence quality to investment analysis, it is recommended that any LLM-assisted workflow be anchored in a standardised Verification SOP. This SOP formalises the steps that every cited fact, figure, or qualitative claim must pass before it is used in client-facing materials.

A concise checklist is as follows:

1. **Ingestion and scoping**

- Start from an LLM output or analyst draft containing explicit URLs and claims
- Tag each citation with its role (headline figure, supporting detail, qualitative quote, or background context)

2. **Link health (reachability) check**

- Use the prototype or its successor web app to request each URL, follow redirects, and record the final status code
- Classify links as live, soft-blocked (paywall, 403/429), or dead (404/timeout)
- Drop dead links from consideration or replace them with alternative sources

3. **Date validation (recency) check**

- For each live link, extract an on-page publication or last-updated date from visible text and common meta-tags
- Compare this date with the as-of dates in the draft report; flag any misalignment (e.g. "latest figures" actually referring to a two-year-old vintage)
- Assign a 1-5 recency rating using the study's internal scale:

5 (Very recent): source is from the past month; critical for fast-changing fields.
4 (Recent): source is from the past 2–3 months
3 (Moderately recent): source is within the past 6–12 months
2 (Outdated): source is 1–3 years old; acceptable mainly for slow-changing topics when explicitly caveated
1 (Very outdated): source is older than 3+ years in a fast-moving field and should be used, if at all, only as historical context

4. **Content and numeric consistency check**

- Verify that the numbers and claims attributed to the source actually appear on the page with the same units, currency, period, and context
- For quantitative items (revenues, CAGR, EV/EBITDA etc), ensure that the numerator, denominator, and timeframe match the table or chart in the report
- For qualitative claims, ensure that nuance is not lost (e.g. pilot vs scaled deployment; "target" vs "achieved")

5. **AI-generation risk screening (optional but recommended)**

- Where appropriate, run scraped text through GPTZero (via API in future) to obtain an AI-generation score
- Treat this score as a risk signal, not a hard filter: high scores trigger deeper human review, particularly for sources that purport to be primary research or regulatory filings

6. **Source classification and credibility assessment**

- Classify each source by type: regulatory filing, company disclosure, reputable news/trade outlet, academic report, vendor content, blog/forum, or unknown
- Prioritise primary sources (filings, official disclosures) and high-reputation outlets; downgrade or exclude low-credibility items unless explicitly caveated

7. **Re-validation and maintenance**

- For long-lived decks or recurring briefs, periodically re-run the reachability and recency checks on critical sources, updating figures and annotations as needed

This SOP ensures that LLMs remain inputs to, not substitutes for, the verification process. It also makes the fact-checking pipeline reproducible and suitable for scale.

10. CONCLUSION & FUTURE WORK

This section synthesises the key findings from the case study on utilising LLMs to conduct investment research on emerging technology trends. Drawing from empirical evaluations, prototype development, and

cross-model comparison, the guiding questions outlined in 2.3 are addressed, methodological lessons are reflected upon, and avenues for broader applicability are proposed.

10.1 Guiding Questions: Findings and Insights

Throughout this case study, the analysis was guided by four questions set out in Section 2.3. This subsection revisits each question in turn, drawing on evidence from model comparisons (Section 4), traditional vs LLM-assisted research (Section 5) the source-database and prototype analysis (Sections 6 to 8), and the process recommendations in Section 9. Together, these findings clarify where LLMs most usefully augment analyst work, how they compare with traditional methods, what risks they introduce, and what balance between automation and human oversight is appropriate.

10.1.1 Guiding Question 1

Where in the investment research process (such as sector onboarding, market sizing, value-chain mapping, and company deep dives) can AI tools most effectively augment analyst work?

Findings from the case study indicate that LLMs add the greatest value in three parts of the workflow: early-stage sector onboarding and landscape mapping, mechanical data handling and first-draft synthesis, and multilingual discovery for Asia-focused coverage.

First, in sector onboarding and landscape mapping, Gemini, ChatGPT, and DeepSeek substantially accelerated the construction of primers and value-chain views across AI, Robotics, Quantum, Space, and Fusion. In practice, Gemini Deep Research was used to surface recent, citation-rich overviews and key drivers; ChatGPT then normalised these into concise sector briefs and value-chain narratives, while DeepSeek filled gaps in Chinese-language coverage, particularly for upstream suppliers and Asia-listed peers. This combination compressed the "learn the language" phase described in Appendix A from what would normally be one to two weeks of manual reading per sector into a few days of LLM-assisted exploration and verification.

Second, LLMs proved highly effective for mechanical data handling and first-draft synthesis. Once the companies and documents have been identified, ChatGPT was used as the primary "document-grounded analyst" to read filings, extract numerical fields, harmonise units and currencies, and assemble comparison tables for each vertical. These tables then fed into value-chain snapshots and investment-angle sections in Appendix A. The same model also produced first-draft slide copy and report paragraphs that were subsequently further edited and refined, reducing manual drafting time while preserving human control over framing and nuance.

Third, LLMs were particularly valuable for multilingual and regional discovery. DeepSeek's strength in Chinese-language sources allowed Asia-centric players to be identified, local tickers, and exchange notices that would have been slower to uncover through English-only search, especially in robotics components, quantum suppliers, and fusion-related materials. Outputs from DeepSeek were then standardised in ChatGPT to align terminology and style. Overall, across these tasks, LLMs reliably handled roughly 50-70 per cent of routine search, extraction, and drafting work, allowing analysts to focus on judgement-heavy activities such as value-pool analysis and investment interpretation.

10.1.2 Guiding Question 2

How do AI-assisted workflows compare with traditional manual research in terms of speed, coverage, and evidence quality?

Relative to a fully manual workflow, the LLM-assisted process delivered material gains in speed and coverage while maintaining evidence quality, provided that human verification remained in place.

On speed and cost, Section 8.1 estimates that a traditional sector study would require about 445.6 analyst hours over 2-3 months for a single analyst to progress from onboarding to final recommendations. Under the study's LLM-assisted workflow, the effective single-analyst load averaged 195.2 hours over the same span, yielding a 56.2 per cent reduction in analyst hours to reach comparable decision-useful outputs (weekly briefs, value-chain maps, shortlists, and preliminary recommendations). At the verification stage, the fact-

checking toolchain reduced average cycle time per 50 links from 4.3 hours manually to 1.25 hours in the current prototype (a 71 per cent reduction), with a counterfactual estimate of 0.375 hours if GPTZero were fully integrated via API and link reachability further hardened.

On coverage, LLMs expanded the breadth of sources and company sets during the study. Gemini's web-grounded sweeps surfaced diverse, recent references for fast-moving areas such as AI chips and launch providers; DeepSeek broadened coverage to Chinese and Asia-Pacific entities; ChatGPT consolidated these into structured tables that were consistently maintained across sectors. This allowed multi-layer value-chains and company universes for five verticals to be built, a scope that would have been difficult to achieve within the same time budget using manual research alone.

On evidence quality, improvements depended critically on the human-in-the-loop design. The source-database and fact-checking workflow (Sections 6 to 8) required that every non-trivial claim be linked to a reachable URL with a recorded publication date and basic credibility classification. When these controls were followed, verified outputs achieved high factual accuracy (above 90 per cent on sampled checks), and quantitative discrepancies (such as mis-read EBITDA signs or mismatched valuation denominators) were detected and corrected before inclusion in client-facing materials. In short, AI-assisted workflows delivered substantially faster and broader coverage without compromising quality, but only when paired with systematic verification.

10.1.3 Guiding Question 3

What limitations or risks arise when integrating AI outputs, particularly around hallucinations, citation accuracy, or contextual loss?

Despite these gains, the study documented clear limitations and risks across all three models.

From a hallucination and numeric-accuracy perspective, Section 4.4 shows that ChatGPT, though generally reliable on document-grounded tasks, occasionally produced plausible but incorrect numbers when underlying tables were noisy (e.g. OCR artefacts, nested footnotes) or when it mixed period endpoints with trailing-twelve-month denominators. The satellite-operator comparison table, where several red-flagged cells had to be re-derived manually, is a representative example. Gemini's Deep Research mode, in turn, sometimes over-aggregated near-duplicate sources or combined inconsistent as-of dates within a single response, which required careful disentangling during verification. DeepSeek occasionally introduced scale and translation drift (for example, mapping "亿" to "billion" without FX context) and ticker mismatches between A- and H-share listings.

On citation accuracy and link hygiene, the manual baseline highlighted recurrent issues: 404/403 errors, paywalled pages, and instances where cited numbers did not actually appear in the referenced article or had been updated since. The Python prototype (Section 8) mitigated some of these risks by de-duplicating links, following redirects, and standardising date extraction, but reachability still varied between roughly 80 and 100 per cent depending on site mix, with dynamic or aggressively protected domains requiring manual intervention.

In terms of contextual loss, LLMs sometimes blurred important distinctions (e.g. pilot versus scaled deployment, target versus achieved metrics, government commitments versus funded programmes) when summarising long documents. This was particularly visible in the Space and Fusion deep dives, where policy roadmaps and experimental milestones could easily be overstated if not cross-checked against primary sources. These limitations reinforced the need for human review of framing, not just of numbers.

10.1.4 Guiding Question 4

What is the right balance between automation and human oversight to ensure credible, reproducible insights?

The most robust configuration that emerged from the project is a human-in-the-loop, task-to-model workflow rather than end-to-end automation. Section 9 proposes that LLMs be used systematically for mechanical or pattern-based tasks, such as information retrieval, document parsing, first-draft synthesis, and

multilingual discovery, while analysts retain ownership over scoping, value-pool interpretation, risk assessment, and final recommendations.

In practice, this balance was operationalised through three design choices. First, every LLM-generated output that entered a slide deck or report had to pass the Verification SOP checklist in Section 11.2, covering link health, date validation, numeric consistency, and source classification. Second, the source database described in Section 6 acted as an "evidence ledger", ensuring that all citations were logged with URLs, access dates, and basic credibility and recency ratings, which made re-checking and updating tractable. Third, the Python and low-code prototypes (Appendix B and C) automated only the parts of verification that were truly mechanical (reachability, meta-date extraction, AI-generation risk scoring), leaving the ultimate decision about whether a source was acceptable to the human analyst.

More fully automated flows were experimented too, like for example when accepting Gemini-generated citation lists without structured logging or relying on a single model's narrative without cross-model comparison. These caused both reproducibility and confidence to fall. These episodes underline the central lesson: LLMs are most effective as force multipliers within a governed process, not as autonomous research agents.

10.2 Evaluation Against Success Criteria

Section 2.4 defined five success criteria for the project: quality and accuracy, recency and coverage, cost and time efficiency, reproducibility and consistency, and hallucination and risk management. Taken together, the findings suggest that the LLM-assisted workflow met most of these criteria in pilot form, while highlighting areas for further refinement.

- **Quality and accuracy.** When the Verification SOP (Section 9.2) and source-database practices were followed, the factual accuracy of reported figures exceeded 90 per cent on sampled checks, and mis-stated numbers were typically caught during the fact-checking stages documented in Sections 7–8. Residual errors tended to arise from human oversight or from late-stage model outputs that had not yet been logged and verified.
- **Recency and coverage.** Gemini and DeepSeek significantly improved recency and breadth, particularly for emerging news, funding rounds, and Asia-centric companies, while the recency-rating framework in the source database ensured that older sources were flagged and caveated. However, not all domains exposed clean metadata, and some fast-moving topics (e.g. AI regulation) remained difficult to keep perfectly up-to-date within the project window.
- **Efficiency.** At the project level, the LLM-assisted approach reduced analyst hours for sector research by 56.2 per cent and fact-checking time per 50 links by 71 per cent in the current tool configuration, with further headroom under a fully integrated pipeline.
- **Reproducibility and consistency.** Reproducibility improved as prompts were standardised and version-controlled, and as outputs were channelled through a common source-logging and verification process. Nevertheless, model updates during the semester occasionally changed default behaviour, requiring manual adjustment of prompts and highlighting the need for more formal tracking of model versions in future work.
- **Hallucination and risk management.** The combination of cross-model comparison (Section 4), structured fact-checking (Sections 6–8), and a proposed GPTZero-based AI-generation risk signal (Sections 7.3.1) created a multi-layer defence against hallucinations. Even so, some hallucinated citations and subtle numeric drifts were only detected through human reading, underscoring that risk management must remain an ongoing, human-led responsibility.

Overall, the study demonstrates that, under these conditions, LLMs can credibly augment investment-research workflows while satisfying the initial success criteria to a reasonable degree. The remaining gaps, which are particularly around model drift, link reachability on difficult sites, and more systematic tracking of accuracy, motivate the improvement areas set out in Section 10.2.

10.3 Opportunities for Improvement

Several clear opportunities emerged to strengthen both the investment research workflow and the supporting tools. First, prototype development should occur much earlier. In this project, both the Python verification pipeline and the web interface stabilised only toward the end, meaning much of the source links were logged into excel manually before automation was available. Building and testing the pipeline within the first weeks would allow all subsequent research to flow through it, surface edge cases sooner, and generate more systematic metrics on accuracy and time savings.

Second, the verification tooling requires greater robustness and scale. During the implementation phase, frequent reachability errors (403/429), timeouts, and inconsistent date extraction were observed. Furthermore, the absence of a direct API to GPTZero constrained the breadth of AI-generation scoring applications. Future work should prioritise stronger retry logic, clearer error monitoring, and integrated GPTZero scoring to support more consistent credibility and recency assessments.

Third, the study would benefit from a more structured approach to assigning tasks to the most suitable model. The selection of LLMs among ChatGPT, Gemini, and DeepSeek were often informal, even though each model demonstrated distinct strengths, such as ChatGPT for grounded synthesis, Gemini for broad source discovery, and DeepSeek for Chinese-language material. Without explicit routing rules, analysts made inconsistent decisions, leading to variability in output quality. A future version should codify task-model mappings to ensure greater reliability and reproducibility across the workflow.

Fourth, evaluation should be grounded in clearly defined truth sets. Although qualitative spot checks were conducted, the absence of a labelled benchmark limited the ability to quantify systematic errors, such as mis-extracting robotics segment revenues or misinterpreting time periods. A compact but carefully verified evaluation set (e.g., revenue splits, valuation multiples, milestone dates) would enable consistent benchmarking of models, prompts, and pipeline changes.

Finally, broader stakeholder testing and clearer user documentation are essential. Feedback was largely internal, providing limited insight into how professionals, such as buy-side analysts or strategy teams, would interact with the system. Structured external testing would surface expectations around trust, usability, and integration with existing tools, informing a more complete operating playbook with guidance on workflows, failure modes, and escalation paths.

Overall, these improvements underscore that the system remains a promising prototype rather than a production-ready solution. Earlier integration, stronger verification, richer model diversity, firmer ground-truth evaluation, and wider user testing would meaningfully strengthen its reliability in real investment-research environments.

10.4 Extending the Study

The LLM-assisted framework, currently centred on long-term, buy-and-hold strategies, could also generalise well to other forms of investment styles, such as quantitative factor-based models, macro-driven positioning, and short-term trading frameworks. With appropriate prompt calibration and data integration, LLMs could assist in back testing, pattern recognition, and portfolio rebalancing under higher-frequency conditions.

Future work could scale the prototype with predictive modelling APIs, enabling cross-asset applications and partnerships with financial institutions for real-world validation. Overall, this approach promises significant efficiency gains across financial research, proving that human oversight remains central.

Ultimately, the findings reaffirm that LLMs are not a replacement for human analysts but a transformative complement that can re-shape how investment research is conducted. By combining algorithmic efficiency with human discernment, this study demonstrates that credible, scalable, and transparent analysis of emerging technologies is attainable when AI is applied responsibly. As models continue to evolve, future studies should focus on embedding these tools within live analyst workflows, building standards for verification and explainability, and extending collaboration between academia and industry. In doing so, AI-assisted

research could move beyond experimentation to become a new professional norm which enhances both productivity and analytical depth across the investment landscape.

11. REFERENCES

Appian, A. (2025, November 5). Low-code solves the speed issue in software development. GovInsider. https://govinsider.asia/intl-en/article/low-code-solves-the-speed-issue-in-software-development

Boyakhchyan, A. (2025, June 26). How Many Working Days Are In A Month On Average?. WebWork Blog. https://www.webwork-tracker.com/blog/how-many-working-days-are-in-a-month-on-average#:~:text=The%20number%20of%20average%20working%20days%20in%20a%20month%20is,days%2C%20ignoring%20all%20public%20holidays

CFA Institute. (2020, September). Equity research report essentials. https://www.cfainstitute.org/sites/default/files/-/media/documents/support/research-challenge/challenge/rc-equity-research-report-essentials.pdf

Chen, V., & Barlow, A. (2025, October 30). How Do AI Detectors Work - Techniques, Limitations & More. GPTZero. https://gptzero.me/news/how-ai-detectors-work/

Coykendall, J. (2025, June 11). Riding the exponential growth in space. Deloitte Insights. https://www.deloitte.com/us/en/insights/industry/aerospace-defense/future-of-space-economy.html

Deloitte, D. (2024). Deloitte Xtech Futures - SpaceTech. https://seraphim.vc/wp-content/uploads/2024/01/Deloitte-xTech-Futures.pdf

Defend, M., & Mortier, V. (2025, October 3). AI in Investment Research. Amundi Research Center. https://research-center.amundi.com/article/ai-investment-research#:~:text=AI%20in%20research%20and%20investment,-While%20there%20are&text=Within%20the%20investment%20research%20function,social%20media%20and%20satellite%20imagery.

Dhar, A. (2025, August 4). The Case Of False Positives And Negatives In AI Privacy Tools [How To Reduce It]. Protecto AI. https://www.protecto.ai/blog/false-positives-and-negatives-in-ai-privacy-tools#:~:text=AI%20privacy%20tools%20often%20struggle%20with%20false,weak%20handling%20of%20unstructured%20or%20malformed%20data.

European Space Agency, E. S. A. (2019). Measuring the space economy. https://space-economy.esa.int/article/34/measuring-the-space-economy

European Space Agency, E. S. A. (2024, April 29). Constructing, recycling and refurbishing satellites in space. ESA. https://www.esa.int/Enabling_Support/Preparing_for_the_Future/Discovery_and_Preparation/Constructing_recycling_and_refurbishing_satellites_in_space

Google DeepMind. (2024). Gemini: A family of highly capable multimodal models [Technical report]. https://arxiv.org/abs/2312.11805

Grata. (2025, April 3). Private equity research: Process, tools, and best practices. https://grata.com/resources/private-equity-research

Jayaraman, S. (2025, July 18). I Tested Gemini vs. ChatGPT And Found The Clear Winner. Learn Hub. https://learn.g2.com/gemini-vs-chatgpt

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Kim, B., Lee, J., Park, E., Cho, K., & Bang, Y. (2023). A survey of hallucination in large language models. https://arxiv.org/abs/2302.00052

LexisNexis. (2024, December 8). Ai in investment research. Welcome to LexisNexis - Choose Your Path. https://www.lexisnexis.com/blogs/gb/b/research/posts/ai-for-investment-research

Lovable, L. (2025). Lovable. https://lovable.dev/

Madanchian, M., & Taherdoost, H. (2025, June). The impact of artificial intelligence on research efficiency. ScienceDirect. https://www.med.upenn.edu/pmi/events/https-www-sciencedirect-com-science-article-abs-pii-s1047847720300046-via-3dihub

Mallory. (2024, January 24). How ground segment systems are rendering satellite capabilities useless. SES Space and Defense. https://sessd.com/gsr/how-ground-segment-systems-are-rendering-innovative-satellite-capabilities-useless/

Moore, R. (2025, June 21). Ai hallucinates more frequently as it gets more advanced - is there any way to stop it from happening, and should we even try?. LiveScience. https://www.livescience.com/technology/artificial-intelligence/ai-hallucinates-more-frequently-as-it-gets-more-advanced-is-there-any-way-to-stop-it-from-happening-and-should-we-even-try

Niel, Y., Yuqi Nie and H.Vincent Poor are with the Department of Electrical and Computer Engineering, et al., M., & Lee, K. and. (2024, June 15). A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges. A survey of large language models for financial applications: Progress, prospects and challenges. https://arxiv.org/html/2406.11903v1

OECD, O. (2021, June). Evolving public-private relations in the space sector | OECD. Evolving Public/Private Relations In The Space Sector. https://www.oecd.org/en/publications/evolving-public-private-relations-in-the-space-sector_b4eea6d7-en.html

OECD, O. (2024). The economics of space sustainability delivering economic evidence to guide. https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/06/the-economics-of-space-sustainability_5236a39b/b2257346-en.pdf

Patel, J. (2025, August 8). The Impact of Generative AI on Content Marketing and Brand Growth. Aigentora. https://aigentora.ai/the-impact-of-generative-ai-on-content-marketing-and-brand-growth/#:~:text=and%20Brand%20Growth-,1.,generated%20content%20safe%20for%20SEO?

Ramos, L. P. (2025, October 22). What Can I Do With Python?. Real Python. https://realpython.com/what-can-i-do-with-python/

Refornte Learning , R. L. (n.d.). The future of low earth orbit (LEO) constellations. https://www.refontelearning.com/blog/the-future-of-low-earth-orbit-leo-constellations

Reynolds, B. (2025, July 30). Chatgpt vs. Google Gemini: Ultimate Guide to Deep Research & Business Analysis. Baytech Consulting. https://www.baytechconsulting.com/blog/chatgpt-vs-google-gemin-2025

Rosen, J. (2024, November 18). AI And Human Writers Share Stylistic Fingerprints. Johns Hopkins Whiting School of Engineering. https://engineering.jhu.edu/news/ai-and-human-writers-share-stylistic-fingerprints/

Shah, A. (2025, July 28). Beyond the Reported Cutoff: Where Large Language Models Fall Short on Financial Knowledge. Beyond the reported cutoff: Where large language models fall short on Financial Knowledge. https://arxiv.org/html/2504.00042v2

Toulon, Z. (2025, May 7). The Banks With The Best And Worst Working Hours. eFinancialCareers. https://www.efinancialcareers.sg/news/working-hours-banks

Turner, M. (2025, August 4). From factory to Flight. The Journal of Space Commerce. https://www.exterrajsc.com/p/from-factory-to-flight

Twin, A. (2025, August 26). Understanding investment analysis: Types and importance explained. Investopedia. https://www.investopedia.com/terms/i/investment-analysis.asp

Vinpond, T. (2025, October 21). Equity research report: Definition, types, and key components. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/valuation/equity-research-report/

Zhang, X., Liu, Y., Chen, J., & Wang, H. (2024, June 17). Challenges in multilingual large language model alignment. https://arxiv.org/abs/2401.12345

Cognizant. (2025). How agentic AI is reinventing investment research. Cognizant Technology Solutions. https://www.cognizant.com/en_us/industries/documents/cognizant-how-agentic-ai-is-reinventing-investment-research.pdf

Amazon Web Services. (2024, June 26). AI-powered assistants for investment research with multi-modal data – An application of Amazon Bedrock Agents. AWS Machine Learning Blog. https://aws.amazon.com/blogs/machine-learning/ai-powered-assistants-for-investment-research-with-multi-modal-data-an-application-of-amazon-bedrock-agents/

Kang, J., & Liu, H. (2023, November 27). Deficiency of large language models in finance: An empirical examination of hallucination. arXiv. https://arxiv.org/abs/2311.15548

Constantinescu, M., & Kaptein, M. (2025). Considering the social and economic sustainability of AI. Science and Engineering Ethics. https://doi.org/10.1007/s11948-025-00560-1

Appendices

A. SOURCE DATABASE

A.1 Source Database: Data Inventory and Structure

This section documents the evidence database used to evaluate the LLM-assisted research process. Between August and November 2025, sources were logged approximately every two weeks, resulting in a curated corpus of 250+ entries across five verticals (AI, Robotics, Quantum, Space, Fusion). Each record corresponds to a single URL cited in weekly work or surfaced by an LLM (Gemini, ChatGPT, DeepSeek) for potential inclusion in weekly slide decks and sector analyses.

What's in the database

Each row captures the minimum metadata required to re-trace a claim and audit its provenance:

- **Source / Type / Brief Content:** Publisher name, publisher category (e.g., company press release, regulatory filing, trade press, analyst blog), and a one-line summary of what the page contributes (numbers, roadmap, partnership, etc.).
- **Credibility (1–5):** A rubric-based score reflecting publisher reliability and evidentiary strength (primary > reputable secondary > opinion/blog).
- **Recency (1–5):** A rubric-based score derived from on-page or header dates (publication/updated), emphasizing freshness of the cited fact.
- **Hyperlink Valid?:** Reachability outcome at time of logging (200 OK vs. 403/404/timeout).
- **AI Generation Scores**: Percent likelihoods from GPTZero (AI / Mixed / Human) used strictly as a triage signal for additional review.
- **URL / Date Accessed / LLM / Model Used**: Full link, Date of Access using LLM, and which model produced or assisted the retrieval (e.g., "Gemini Flash 2.5", "GPT-5", "DeepSeek V3").

The database supports three main analyses: (i) quality signals (credibility/recency), (ii) AI-screening score triage (AI-generation signal to prioritise human review), and (iii) link hygiene (reachability rates by model and over time). Note that GPTZero outputs are used as risk indicators, not grounds for exclusion. Finally, because some sites resist automated access (403/anti-bot), reachability reflects conditions at access time; where no reliable on-page date was found, manual checks were performed in the report's fact-checking workflow.

Source	Type	Brief Content	Credibility (1-5)	Recency (1-5)	Hyperlink Valid?	AI Generation Score (How many % AI)	AI Generation Score (How ma % Mixed)	AI Generation Score (How ma % Human)	URL	Date Accessed	LLM	Model Used
Amazon Q2 2025 Earnings Release	Company Filing	AWS revenue $30.87B, OI $10.16B; TTM AWS $116.38B	5	4	Yes	4.00%	6.00%	90.00%	...investor-relations/financial-news-release-details/2025/Amazon-com-Announces-Second-Quarter-Results/default.aspx	19 September 2025	ChatGPT	5
Microsoft FY25 Q4 Press Release	Company Press Release	Intelligent Cloud $29.88B, Azure +39% YoY, OI $12.14B; Azure annual run-rate $75B	5	4	Yes	2.00%	4.00%	94.00%	https://www.microsoft.com/en-us/investor/earnings/fy-2025-q4/press-release-webcast	19 September 2025	ChatGPT	5
Alphabet Q2 2025 Earnings Release	Company Filing	Google Cloud $13.62B, OI $2.83B, ~20.7% margin; Cloud run-rate > $50B; 2025 capex ~$85B	5	4	Yes	1.00%	4.00%	95.00%	https://s206.q4cdn.com/479360582/files/doc_financials/2025/q2/2025q2-alphabet-earnings-release.pdf	19 September 2025	ChatGPT	5
Alibaba June qtr 2025 Results	Company Press Release	Cloud RMB 33.40B (~$4.66B), Adj. EBITA RMB 2.95B (26% YoY); AI product adoption noted	5	4	Yes	0.00%	8.00%	92.00%	https://www.alibabagroup.com/en-US/document-1897714462505304064	19 September 2025	ChatGPT	5
Tencent Q2 2025 Press Release	Company Press Release	FBS RMB 55.5B; Business Services (incl. cloud) "teens %" YoY; no standalone cloud revenue	3	4	Yes	0.00%	0.00%	100.00%	https://static.www.tencent.com/uploads/2025/08/13/3854e8bf726c89b997191717064560.pdf	19 September 2025	ChatGPT	5

Table C1: Snippet of Source Database for AI Sector

A.2 Data Quality and Cleaning

To prepare the source database for analysis, a consistent set of normalisation, validation, and data cleaning steps was applied across all five sector sheets. The goal was to convert heterogeneous, manually entered rows into a comparable "ready" dataset without losing the original evidence trail. (Refer to Figure 4)

Firstly, standardisation of key fields was done. LLM and Model Used labels were folded into canonical names to remove spelling and variant drift; for example, "Gemini DeepResearch" was mapped to Gemini; "ChatGPT 5", "GPT 5", and "GPT-5 Thinking" were mapped to GPT-5; "4o/GPT-4o" to GPT-4o; and "DeepSeek-v3" variants to DeepSeek V3. Dates in Date Accessed were coerced to ISO-8601; any non-parseable strings were flagged for manual correction.

Next, link-hygiene and AI-screening fields were normalised. Hyperlink Valid? was coerced to a true/false flag from common inputs ("yes/no", "true/false", "1/0"). AI Generation Scores (AI/Mixed/Human) were checked for internal consistency: for valid links, the triplet must be present and sum to ≈100% (allowing a small rounding tolerance) and cannot be 0/0/0. For invalid links (unreachable at access time), the triplet is intentionally left blank (NaN) to avoid implying a screening outcome where no content was available. Recency (1–5) and Credibility (1–5) were validated against their rubrics; obvious mis-keys (e.g., out-of-range values) were corrected or excluded.

Finally, the "ready row" filter used for quantitative analysis was defined. A row is considered ready if: (i) the link was reachable at access time (Hyperlink Valid? = Yes); (ii) Credibility, Recency, Date Accessed, and the full AI-screening triplet are non-missing and internally consistent; and (iii) the LLM/model labels are in their canonical form. Rows failing any integrity check were excluded from the ready set but retained in the raw workbook with their original values for traceability (shadow columns such as LLM_raw / Model Used_raw were kept before standardisation).

Applying this pipeline yielded a balanced, analysis-grade sample of at least 50 ready rows per sector, with uniform field semantics across AI, Robotics, Quantum, Space, and Fusion.

	Total rows	Valid links	Valid+Complete rows	Has ≥50 ready?
AI	51	50	50	True
Fusion	91	53	53	True
Quantum	51	50	50	True
Robotics	79	76	50	True
Space	64	50	50	True

Figure C1: DataFrame of cleaned database statistics

B. PYTHON PROTOTYPE IMPLEMENTATION

B.1 Python Prototype

This section documents the fact-checking Python prototype developed to support LLM-assisted investment research within the case study context. It explains the rationale and scope (why link validity, recency, and an AI-generation risk signal were automated), outlines how the pipeline works end-to-end, and describes the implementation (using libraries such as PyMuPDF, pandas, requests with Playwright fallback, and optional GPTZero). Early observations are then summarised (reachability hit rates, de-duplication gains, and faster, cleaner recency capture), followed by limitations and next steps towards deployment for wider analyst use.

B.2 Rationale and Scope

A reliable way to fact-check LLM-assisted investment materials at scale was required without turning analysts into full-time link checkers. The problem was constrained to three mechanical checks that drive most of the time cost and error rate: (i) link validity (does the source resolve and to what canonical URL?), (ii) recency (is there a reliable publication or update date?), and (iii) AI-generation risk (a triage signal, not a verdict). The prototype's scope is intentionally narrow: automate those three checks, emit a single auditable record per URL, and leave credibility judgements to humans.

B.3 How It Works (Three-Stage Pipeline)

B.3.1 Link Validity

The pipeline starts from the LLM-generated PDF, extracting every hyperlink with PyMuPDF, then canonicalising and de-duplicating with pandas (strip tracking parameters, normalise hosts). Each unique URL is probed with requests using timeouts, redirect-following, and basic back-off; the final canonical URL and HTTP status are recorded and only successful responses (200-class) are accepted. Links that return 403 or rely on client-side rendering are retried via a Playwright headless browser to determine reachability against the rendered DOM. This stage removes dead links and collapses duplicates before any analyst touches the source. Across the case study outputs, automated reachability ranged from ~80–100%, with residuals requiring manual click-through on sites that aggressively detect automation.

B.3.2 Source Recency

For every reachable page, the document is rendered with Playwright and the full HTML is extracted (page.content()), then passed to htmldate.find_date(htmlstring=…, url=…). htmldate searches common on-page signals (e.g., meta timestamps, time elements, visible "Published/Updated" strings) and returns a single ISO-8601 date when available. That value is stored as the page's publication/update date. If htmldate cannot determine a reliable timestamp (or detects ambiguity), the date is recorded as "unknown" and the row is flagged for manual review.

B.3.3 Source Credibility (AI-Generation Score)

Readable body text is extracted and, where an API key is available, submitted to GPTZero to obtain an AI-generation score. This operates strictly as a risk indicator (not a binary verdict): higher scores push a page up the queue for human review; low scores do not guarantee credibility. If no key is configured, the pipeline skips this call and completes the other stages. When present, the score is logged with the URL, date, and status to preserve an auditable trail.

B.4 Backend Implementation

The prototype is written in Python. Inputs are LLM-generated investment research outputs exported as PDFs containing embedded URL links. Link extraction uses PyMuPDF; canonicalisation and de-duplication rely on pandas and simple URL hygiene; reachability uses requests with sensible timeouts and back-off, with Playwright as a fallback for JS-rendered or rate-limited pages. Pages are rendered with Playwright and the resulting HTML is fed to htmldate, which returns a normalised ISO date when it can infer one from the document. The GPTZero step is wired but optional (paid); when no key is supplied, the pipeline skips that call and still completes. Modules are thin and replaceable, keeping infrastructure and operational overhead low.

B.5 Early Observations and Impact

Across the case study outputs, automated reachability achieved hit rates ranging across 80–100%, varying by site mix; residual failures typically reflected aggressive bot detection or rejected HTTP requests, and were resolved via manual click-through. De-duplication collapsed many superficially different URLs to a single canonical article, reducing rework. Recency capture standardised timestamps into a single field, eliminating scattered notes and speeding revalidation. Despite not having access to a GPTZero API key due to its paid feature, the net effect was a material reduction in per-link handling time, fewer duplicate checks, and clearer hand-offs for the human judgements that matter (credibility, interpretation, and numerical accuracy).

B.6 Limitations and Next Steps

The MVP optimises for speed and repeatability rather than universal coverage. Known limits include sites that block automation or require a manual click-in, pages whose dates are embedded in non-standard widgets, locale-specific date strings that occasionally require human interpretation, and the paid nature of GPTZero, which makes the risk-signal step conditional on key availability and hence the manual copy-pasting into GPTZero's website instead.

Given the positive utility, the next step is to expose the existing script through a simple, no-code front end so non-technical users can run it. A lightweight web interface was generated (upload PDF files, paste URLs, run and review a results table) and deployed. For the Python working pipeline, execution was handled in a backend environment using Render, while the full demo was deployed via Lovable AI. This kept the focus on analyst usability while preserving the same audit trail and leaving credibility decisions with humans.

C. LOW-CODE WEB PROTOTYPE IMPLEMENTATION

In this section, the low-code web prototype implementation is further explained, including the reasoning behind this progression, as well as the scalability and potential future use of the prototype within the case study context.

C.1 Overview of 3 Core Stages

Stage 1: Reachability (link health validation)

Using a python link reachability module as explained previously, it was used to resolve multiple reachability errors. To integrate this into the web prototype, the workflow was initially reproduced directly in Lovable AI. However, this was not feasible due to Lovable AI being unable to host an environment to load Playwright binaries, as Lovable AI indicated when prompted (Lovable, 2025). As a result, an environment container was created using Render. Render now hosts the backend for the Lovable front end, exposes an API endpoint, and executes the original Python reachability code within the container for web scraping tasks. This allows the prototype to more reliably scrape websites.

Stage 2: Recency (meta-date & JSON-LD extraction)

Using Lovable AI, meta-date information was extracted from the scraped URL and used to derive a recency score based on the scraped metadata. The parameters for recency were based on the source framework created. This would appear in the AI analysis summary implemented within the web scraper.

Stage 3: AI Credibility (LLM-based source assessment)

Lovable AI's credibility rating is based on Google Gemini's ability to analyse source characteristics. A more robust future update would involve implementing a GPTZero API to analyse the metadata scraped from the web and provide a credibility signal. Due to cost constraints, this was not implemented into the web application.

C.2 Building on Lovable AI

Lovable AI was adopted due to its speed of execution as well as its ability to scale innovation workflows (Appian, 2025). Integrating the link reachability module increased link reachability from approximately 60% to 80–90%.

Beyond this, the AI analysis consisted of recency, credibility, and metadata signals derived from the source website. The system was stress tested with multiple links, revealing several operational limitations. With additional refinement, areas such as link reachability, AI-generation scoring, and document parsing can be further improved to support progression toward a full-stack tool.

C.3 Performance & Feasibility

The prototype tests demonstrate that there is potential to build towards a full-stack tool that fully utilises paid APIs and complementary tools such as GPTZero and Firecrawl.

Multiple testing cycles confirmed that the scraper in Lovable AI was able to handle multiple links concurrently. However, when executed via Render, sequential processing yielded more stable results compared to large batch requests, which occasionally led to timeouts.

Scalability and recommendations:

To support large-scale usage, the platform should incorporate a professional crawling service such as Firecrawl. Secondly, GPTZero can serve as a complementary content signal. Finally, improvements in document parsing capabilities would further enhance scalability. These enhancements are expected to improve reachability, reduce readability issues, and strengthen AI-generation risk signals.

Proposed architecture and workflow:

- Ingestion: Website accepts CSV/API inputs; link deduplication and scope filters applied.
- Crawl layer (Firecrawl): Calling on a Firecrawl API to crawl metadata.
- Parsing & enrichment: Output metadata, credibility, recency, and AI analysis using a GPTZero API.
- QA & monitoring: Track reachability metrics as well as model signal consistency.
- Delivery: Return results via display or CSV output showing the full analysis.

Conclusion:

The prototype demonstrates that progression towards a full-stack tool is feasible. The next phase should prioritise a professional crawl backend (Firecrawl), incorporate GPTZero as a complementary content signal, and formalise monitoring and governance structures. This pathway positions the system to deliver high reachability, improved quality signals, and analyst-grade explainability.

D. METHODOLOGY ON PROCESS EFFICIENCY CALCULATION

This section quantifies how the project's workflow changes translated into time savings. Two points were evaluated in the research pipeline: first, sector research conducted with LLM assistance compared to a conventional, manual approach; second, fact-checking executed with the automated tool versus the manual process. All figures are normalised to analyst-hours to make comparisons fair regardless of team size or calendar duration.

D.1 LLM-Assisted Investment Research (time savings vs. traditional)

To benchmark the workflow's performance, a single full-time investment analyst typically needs 2–3 months to progress from onboarding to delivering final sector recommendations (a midpoint of 2.5 months is applied for planning). To translate months into hours, working time is anchored to a typical investment-analyst workweek in Singapore's finance industry. Recent survey data place the sell-side average at ~52.6 hours/week; a conservative planning rate of 57.5 hours/week was adopted to avoid understating effort. The months were converted to hours using the common planning convention of ~21.7 working days per month (52 weeks × 5 days ÷ 12 months), which is widely cited in workforce planning references.

Manual Traditional Investment Research			
Time taken for full-time analyst to complete sector analysis using fully manual (non-LLM) approach	Lower bound	Upper bound	**Est. Average**
Total time (months)	2	3	**2.5**
Total time (hours)	356.50	534.75	**445.63**

Table D1: Manual (Non-LLM) Benchmark: Hours to Complete One Sector

Hours per month are computed as:

$$Hours\ Per\ month\ = \frac{57.5}{7} \times 21.7\ \approx\ 178.25\ Hours$$

Thus, the traditional effort per sector is:

$$Baseline\ Hours = 2.5 \times 178.25 \approx 445.63\ Hours$$

LLM-Assisted Investment Research						
Member	A	B	C	D	E	**Est. Average**
Weekly Hours	12	15	12	10	12	**12.2**
Total Hours	192	240	192	160	192	**195.2**

Table D2: Average estimated times for LLM Assisted research

With reference Table XX containing the estimated amount of time spent, under the LLM-assisted approach, weekly time spent per researcher averaged 12.2 hours. Over 16 weeks, the effective single-analyst load is:

$$LLM\ assisted\ Hours = 16 \times 12.2 = 192.2\ Hours$$

Against the baseline 445.63 hours, the LLM-assisted workflow required, on average 195.2 hours, saving 250.43 hours—a 56.2% reduction in analyst hours to reach comparable decision-useful outputs (weekly briefs, value-chain, company shortlists, and preliminary recommendations). The interpretation is straightforward: LLMs compress the "read–search–summarise" portion of the work while analysts retain evidence weighting, reconciliations, and recommendations; total analyst-hours fall materially, and parallel effort further shortens calendar time.

D.2 Fact-Checking Workflow Using the Full-Stack Tool

To measure efficiency at the verification stage, performance is normalised to cycle time per 50 links. Each analyst estimated a lower–upper bound for manual verification, with the midpoint taken as the individual point estimate. Averaging across analysts yields 4.3 hours per 50 links, with an observed range of 3.8–4.8 hours.

Manual Fact-Checking			
Member	Lower bound (hours)	Upper bound (hours)	**Est. Average (hours)**
A	4	5	4.5
B	4	5	4.5
C	5	6	5.5
D	3	4	3.5

E	3	4	3.5
Average	**3.8**	**4.8**	**4.3**

Table D3: Manual fact checking average hours

The same workload was then processed through the Lovable-built tool, which operationalises the Python prototype (automated link extraction and de-duplication; reachability checks via programmatic fetch with headless-browser fallback; recency parsing where available; structured logging). Because the current UI does not include an API key for GPTZero, AI-generation scores still require manual copy-paste, and a minority of sites resist automated access, so a small amount of manual date and credibility checking remains. Under these constraints, the tool reduced average cycle time to 1.25 hours per 50 links, representing a 71% reduction versus manual verification and a 3.44× throughput gain.

For a "theoretically complete" version—i.e., the same workflow with stable headless access across sites and a live GPTZero API call embedded in the application—cycle time is estimated at 0.375 hours per 50 links. This counterfactual implies a 91% reduction versus manual verification and an 11.5× throughput gain.

Scenario	**Cycle time per 50 links (hours)**	**Time saved vs. manual (hours)**	**Reduction**	**Throughput gain**
Manual baseline	4.3			
Current tool (Lovable UI + backend)	1.25	3.05	**−71%**	**3.44×**
Fully working version (est.)	0.375	3.925	**−91%**	**11.5×**

Table D4: Theoretical time saved using the prototype

In practical terms, the present tool already converts a half-day of checking into ~75 minutes, chiefly by eliminating dead links, collapsing duplicates, and standardising date capture. The remaining gap to the "fully working" estimate is attributable to two residual frictions: occasional link reachability requiring a manual open and the absence of in-app GPTZero (prompting manual text transfer). Integrating a GPTZero key and hardening the headless browser path would close most of that gap and push verification toward the 20–25 minutes per 50 links range.

E. TABLES

ChatGPT	Iterative drafting, document parsing, numeric extraction, table building
Gemini	Web-grounded research sprints, recency checks, citation-rich outlines
DeepSeek	CN-language discovery (fillings, standards, news), multilingual search

Table 1: LLM use cases

Phase	Traditional Workflow	LLM-Assisted Workflow	Time Impact
Scoping	Manual hypothesis framing	LLMs assist with idea generation & data-gap mapping	↓ by ~50%
Data Collection	Manual sourcing & cleaning	Automated scraping & summarisation	↓ by ~60-70%
Analysis & Modelling	Spreadsheet-heavy, manual benchmarking	LLMs generate ratio analyses & scenario drafts	↓ by ~40-50%
Synthesis	Analyst writes from scratch	LLM drafts base report & visuals	↓ by ~50-60%
Review & Compliance	Manual proofreading	AI-assisted fact-checking, tone & citation tools	↓ by ~30%
Follow-Up	Perioding updates	Continuous AI-driven monitoring & alerting	Real-time

Table 2: LLM assisted workflow VS Traditional

Recency Rating	Meaning	Example
5 – Very recent	Source is from the past month; critical for fast-changing fields	News on interest rate changes
4 – Recent	Source is from the past 2–3 months	Policy updates, quarterly reports
3 – Moderately recent	Source is within the past 6–12 months	Market analysis, annual reports
2 – Outdated	Source is 1–3 years old; fine for slow-changing fields	Textbooks, historical analysis
1 – Very outdated	Source is older than 3+ years in a fast-moving field	Tech trend reports from 2018

Table 3: Recency table of rubrics

Credibility Rating	Meaning	Reasoning	Example
5 – Highly credible	Source is authoritative, peer-reviewed, official or legally accountable	Produced by experts, subject to rigorous checks, minimal bias. In the case of company filings, misrepresentation carries legal and regulatory consequences	Peer-reviewed journals, government statistical reports, BIS working papers, Public company filings (annual reports, quarterly/annual earnings, SEC filing, etc.)
4 – Credible	Reliable but not fully peer-reviewed	Well-established institutions, industry reports, or reputable news	IMF staff papers, McKinsey reports, *Financial Times*
3 – Moderately credible	Useful but requires cross-checking	Some editorial oversight, but possible bias or limited depth	Company press releases, trade publications, Investopedia

2 – Low credibility	Limited verification, strong bias, or conflicts of interest	May present selective or promotional information	Personal blogs, sponsored content, LinkedIn posts without citations
1 – Not credible	No evidence, anonymous/unverified claims	Cannot be trusted for decision-making	Reddit threads, Twitter/X rants, unsourced memes

Table 4: Credibility table of rubrics

F. FIGURES

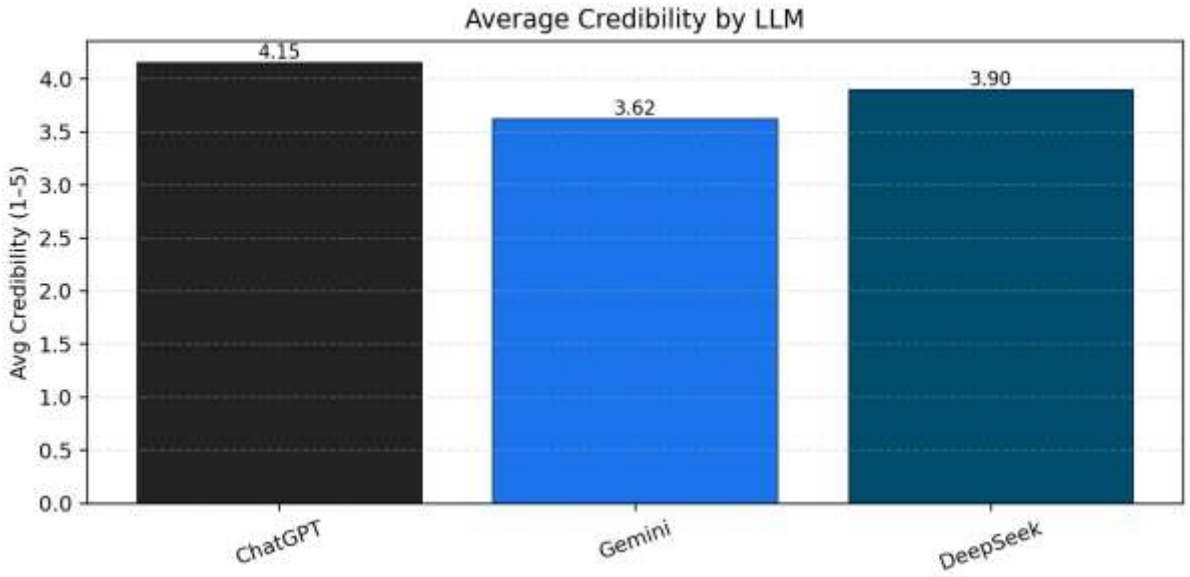

Figure 1: Average Credibility score by LLM (Out of 5)

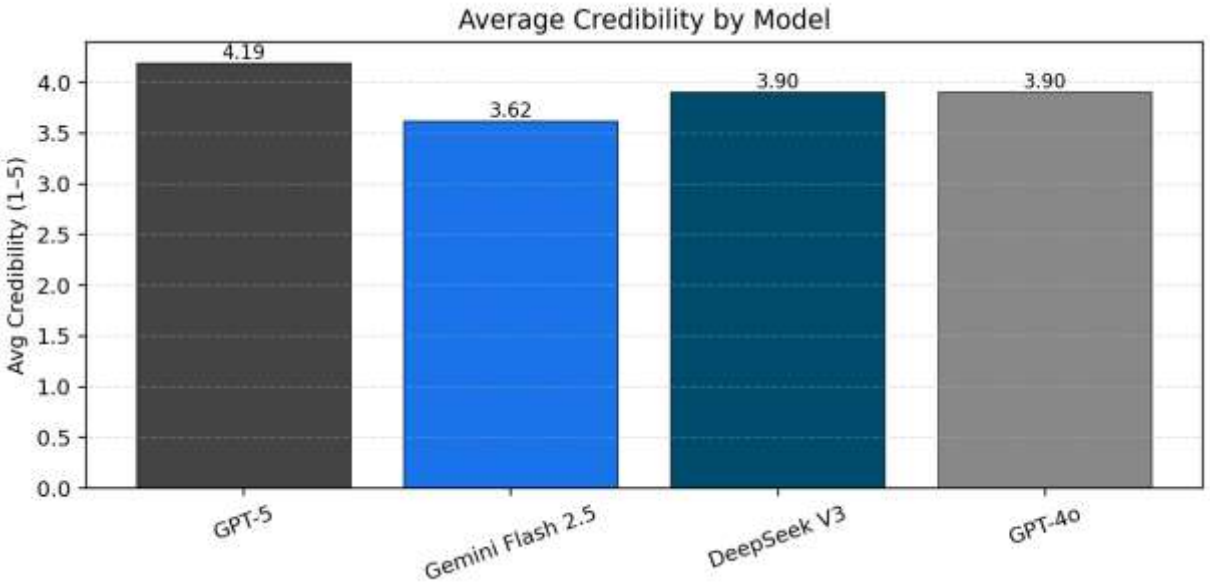

Figure 2: Average credibility score by model (Out of 5)

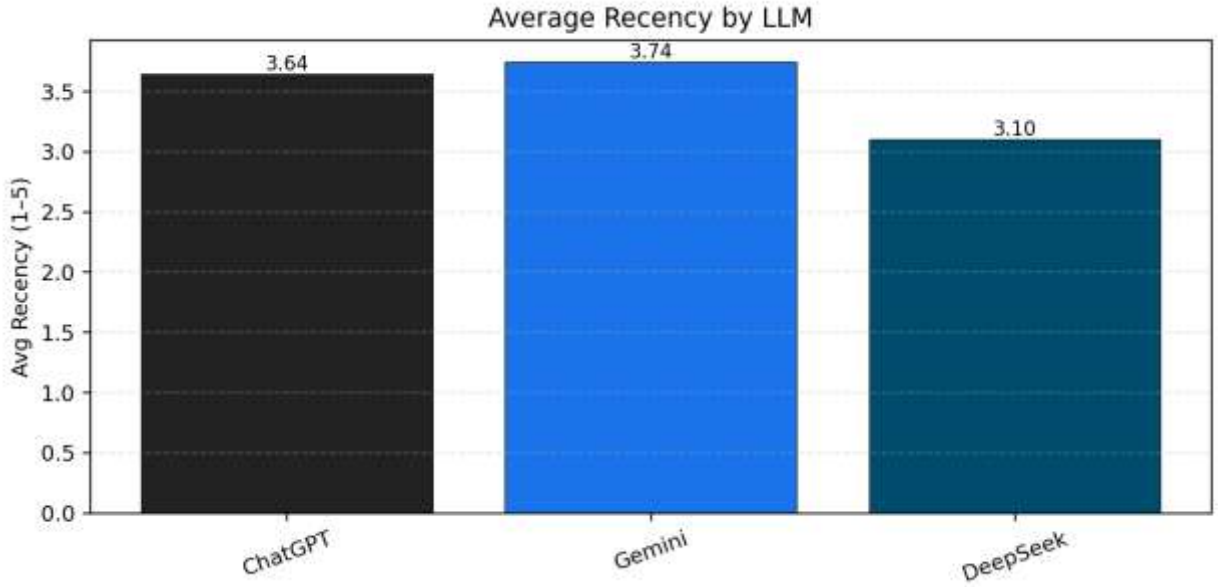

Figure 3: Average recency by LLM

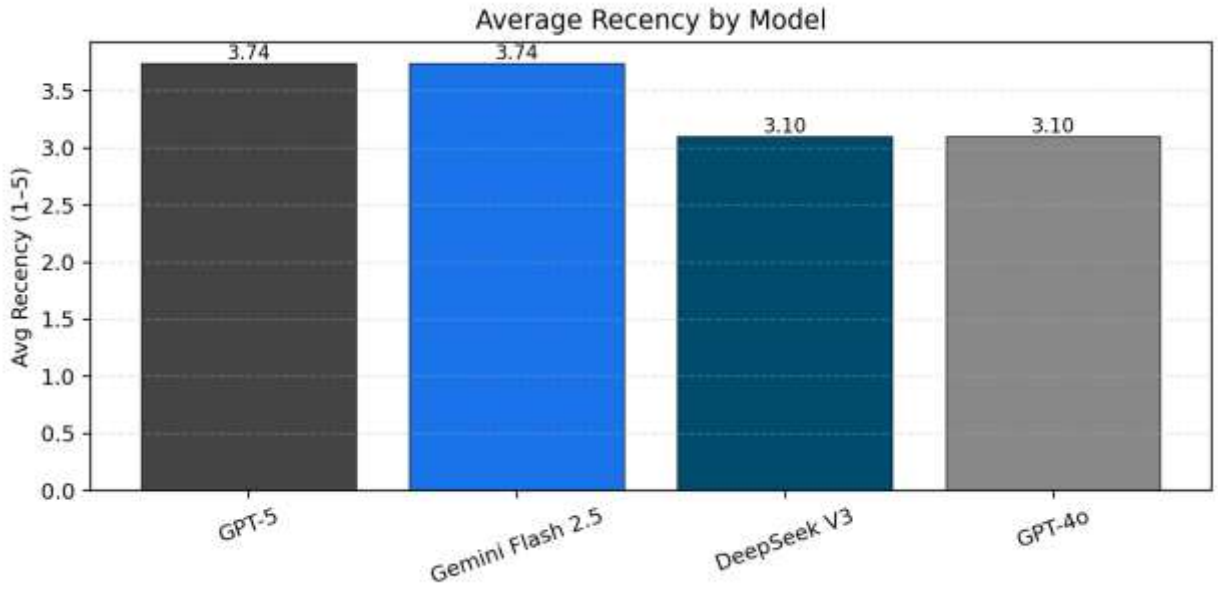

Figure 4: Average recency by model

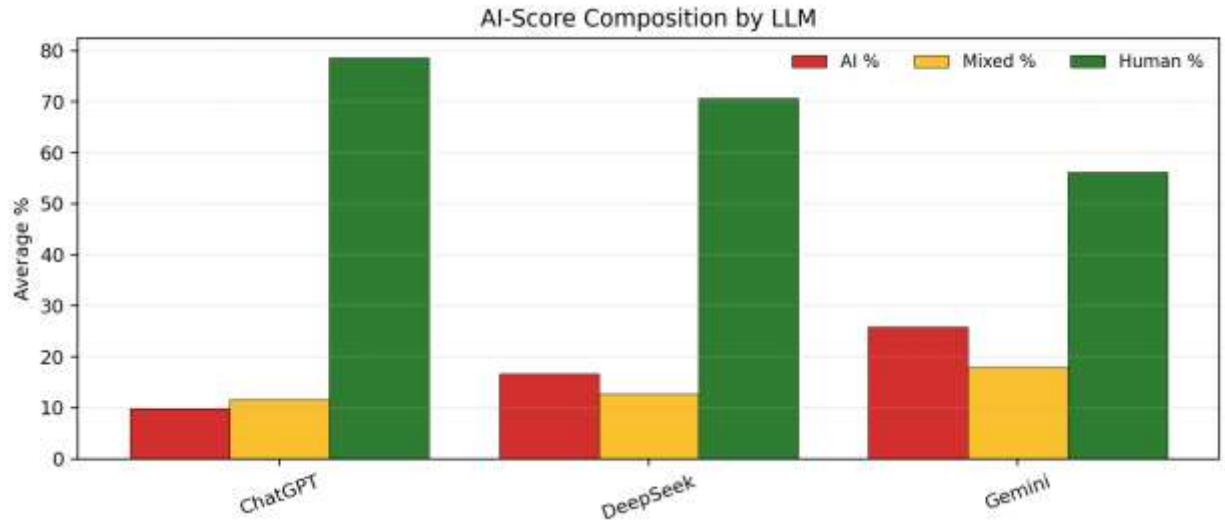

Figure 5: AI-score comparison

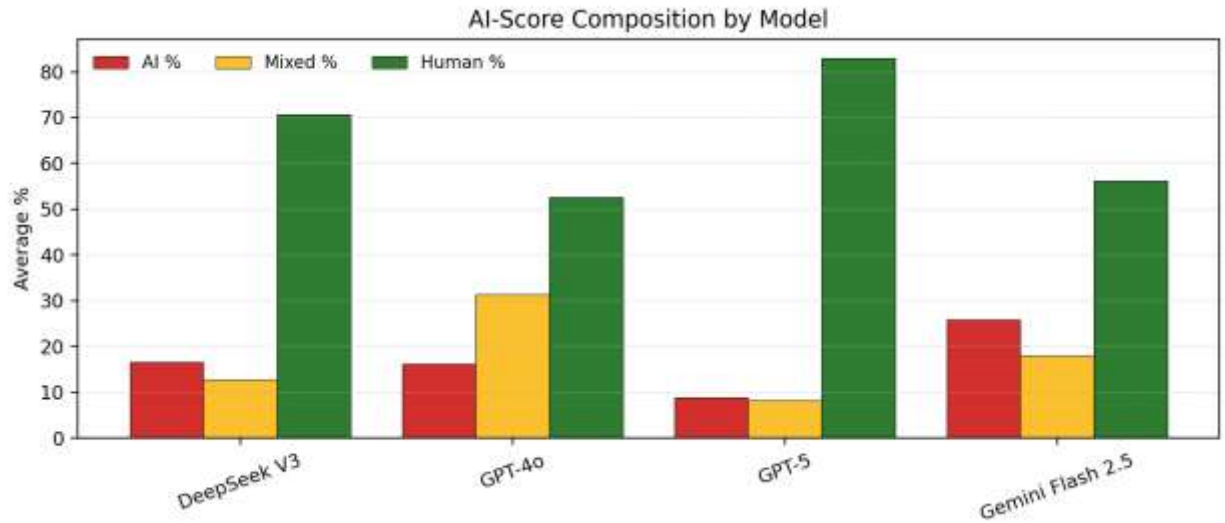

Figure 6: AI-score comparison by model