

Peer Review

Review of: "Agentic Systems: A Guide to Transforming Industries with Vertical AI Agents"

Evgenii Arsentev¹

1. Independent researcher

Summary

The article argues that traditional SaaS and context-aware systems cannot serve domain-specific needs (Sections 1.1 and 1.2), that "Vertical AI agents" close that gap (Section 2), and that such agents are built from four modules — Memory, Reasoning Engine, Cognitive Skills, Tools (Sections 3.3.1 through 3.3.4) — of which Cognitive Skills is claimed as the author's own contribution (Abstract; Section 3.1: "We introduce a new module to the core building blocks of the LLM agent"). Section 4 then sorts agentic systems into three categories — Task-Specific, Multi-Agent, Human-Augmented (Section 4.3) — and walks through one reference pattern for each: the RAG Agent Router (Figure 3), the RAG Orchestrated Multi-Agent System (Figure 4), and the HITL Agent (Figure 5). Everything that holds this up is definitional and illustrative. The article contains no data, no experiment, no benchmark, no deployment record, and no comparison table; the support is a stated architecture plus 14 references, 6 of which (references 4 through 9) are vendor blog posts and product pages. As a teaching map of the current design vocabulary, it reads well, and the figures are sensibly chosen. As a review article in the scholarly sense — a survey with a method, a scope, and an evidentiary standard — it is not yet there, and that is the gap this report is mostly about.

Does the claim hold on the data

There is no data, which is legitimate for a review, so the question becomes whether the claims are of a kind a review can make without any. Several are not. Section 2.1.1 asserts that vertical agents perform "legal contract analysis, medical imaging interpretation, or financial risk assessment, with exceptional precision"; Section 2.1.3 asserts "Faster Turnaround Times: Analyzing, deciding, and executing tasks

within seconds" and "Cost Optimization ... reducing operational expenses"; Section 2 opens with "unprecedented effectiveness"; Section 5.1 concludes that agentic systems "have demonstrated their ability to revolutionize industries, driving operational efficiency." Those are empirical claims about outcomes — precision, latency, cost, error rates — and they are exactly the claims a review is supposed to source. Section 2 carries no citation at all. Neither does Section 5. In the whole article, the citation markers fall in five places: reference 1 in Section 1, references 2 and 3 in Section 3.2, references 4 through 9 in Section 4.2.1, references 10 through 12 in Section 4.2.2, and references 13 and 14 in Section 4.3.1. The entire benefits argument and the entire conclusion are unsourced. That is the single biggest structural problem in the paper: the parts a reader would most want to check are the parts with nothing to check against.

The second problem is the boundary between what vendors say about their own products and what is established. Section 4.2.1 describes AutoGen and Semantic Kernel, the OpenAI Assistants API, Vertex AI Agent Builder, the AWS suite, Anthropic's patterns, and LangChain — each sourced only to that vendor's own blog or documentation (references 4 through 9). The article repeats the marketing framing without marking it as such: AWS "provides a robust suite of tools," Semantic Kernel integrates AI "with a focus on security and scalability." Worse in the other direction, the LangChain entry makes a negative performance claim — "faces challenges with speed limitations when managing complex interactions between multiple agents and tools" — and cites reference 9, the LangChain v0.1 agents documentation, which is not a place where such a finding could come from. Either the claim has a measurement behind it, in which case cite it, or it is the author's impression, in which case say so.

Third, the novelty claim. Cognitive Skills is announced in the Abstract and Section 3.1 as a new core building block, and Section 3.3.3 defines it as a module that "acts as a model hub, equipping the agent with purpose-built models." An LLM that dispatches to a hub of specialist models is prior art with a literature — HuggingGPT and the model-routing line of work, and on the invocation side Toolformer and the tool-use taxonomies. None of it is discussed or cited. Without a related-work paragraph that says what Cognitive Skills has that a typed tool registry does not, the contribution reads as a renaming. The article also does not supply a rule that separates the two buckets it created. RAG sits under Tools (Section 3.3.4), OCR and guardrail classifiers under Cognitive Skills (Section 3.3.3), yet all three are called the same way: the agent sends an input to an external model or service and gets a result back. In Figure 4, the arbitrariness is visible in one frame — LLM Agent 3's Search APIs are labeled Tools, LLM Agent 4's

Guardrail Classifiers are labeled Cognitive Skills (Section 4.3.2). Give a practitioner a new capability, say a fine-tuned entity linker, and this article offers no criterion for which box it goes in.

Fourth, the taxonomy of Section 4.3 is not a partition. Task-Specific Agent, Multi-Agent System, and Human-Augmented Agent are cut along three different axes: scope, cardinality, and who signs off. A RAG Orchestrated Multi-Agent System (Section 4.3.2) with a compliance officer reviewing outputs is both the second and the third category at once — and the article's own fraud-detection example in Section 4.3.3 is precisely that. Three categories chosen for exposition are fine; three categories presented as "three primary types based on their structure, scope, and interaction dynamics" need the categories to be disjoint, or need to say plainly that they are overlapping lenses.

Fifth, and this one is about labeling rather than truth: every block titled "Practical Use Cases" (Sections 4.3.1, 4.3.2, 4.3.3) is a hypothetical. Scenario, then Solution, then the expected outcome, with no organization, no system, no volume, no latency, no accuracy, no before-and-after. Healthcare Decision Support (Section 4.3.1), Healthcare Assistance (Section 4.3.2), and Healthcare Diagnostics and Treatment Planning (Section 4.3.3) are three worked examples in a regulated domain that assume an agent can route to EHRs, diagnostic databases, and clinical guidelines and compile a treatment recommendation. Presented as illustration, that is acceptable. Presented as "practical use cases," it will be read as fielded practice, and it is not.

Reproducibility

For a review article, the reproducibility question is whether a second person, given the stated method, would assemble a similar body of work. Here, there is no stated method at all. The article does not name a search database, a query string, a date window, an inclusion or exclusion criterion, a count of candidates screened, or a reason why these 14 references and not others. Section 4.2 says "At the time of writing this article, there are no universally accepted design patterns" — a survey-shaped claim about the state of a field, and the kind of claim that is only credible if the reader can see what was searched. As it stands, references 4 through 9 read as the six vendors a practitioner in late 2024 would already have had open in browser tabs, and references 10 through 12 as three arXiv preprints from the preceding two months. That may be a fair sample; the reader cannot tell, and neither could the author on a re-run.

The snapshot problem is real here and partly handled. References 4, 5, 6, 7, and 9 carry "Accessed: December 26, 2024," which is correct, and reference 9 is pinned to LangChain v0.1 documentation, which is better. Reference 8 (Anthropic, "Building effective agents") carries a month, November 2024, but no

access date. Vendor agent products in this period changed on a scale of weeks — capabilities described in Section 4.2.1 from an OpenAI DevDay 2023 announcement (reference 5) were already several product generations stale at submission. A version-and-date column in a comparison table would fix this and would also give the article the artifact it currently lacks: there is not a single table in the paper.

The other reproducibility gap is that nothing in the article is executable or checkable by a reader who wants to build what is described. Figures 1 through 5 are architecture diagrams; in the version supplied to me, only their captions came through, so I judge them on what the text says they show and do not attribute content to them. But even from the text, it is clear there is no interface specification, no message schema between orchestrator and sub-agents, no routing policy, no failure or fallback behavior, no cost model, and no pseudocode. The Abstract promises "a level of standardization for Vertical AI agent design patterns." Standardization means something another team can conform to and be checked against. Prose plus a block diagram does not reach that bar.

Finally, a review of agentic systems that never mentions how any of this is measured is missing a limb. There is no discussion of agent evaluation anywhere in the paper — no benchmark, no task suite, no success criterion, no mention of the fact that multi-step agents are notoriously unstable across seeds and model snapshots and that a pattern that works on one model version often does not survive the next. The Cognitive Skills argument in Section 3.3.3 — that purpose-built models beat fine-tuned LLMs on tasks like OCR — is stated as fact and is the one claim in the article that would be straightforward to support with a measurement or a citation, and it has neither.

Specific points

1. **Major — no review method (Sections 1 through 5, throughout).** The article is typed REVIEW ARTICLE but states no search source, no coverage period, no inclusion or exclusion criteria, and no screening counts. Add a short methods subsection: where you searched (arXiv, ACL Anthology, vendor documentation), the window (for example, January 2023 to December 2024), what qualified a framework or paper for inclusion, and what was seen and dropped. Without it, the reader cannot distinguish a survey from a reading list, and the state-of-the-field claim in Section 4.2 has nothing under it.
2. **Major — the benefits argument is entirely uncited (Section 2, including 2.1.1, 2.1.2, 2.1.3).** Not one citation marker appears between the start of Section 2 and the start of Section 3. The claims there are empirical: "exceptional precision" (2.1.1), "reducing errors and manual effort" (2.1.1), "within

seconds" (2.1.3), "reducing operational expenses" (2.1.3). Either attach sources to each or rewrite them as expectations and say so in the text. The same applies to Section 5.1's "have demonstrated their ability to revolutionize industries," which is a results claim in a paper with no results.

3. Major — vendor copy presented as findings (Section 4.2.1, all six entries, references 4 through 9).

Every entry is sourced to the vendor describing itself. Mark it: a sentence at the head of 4.2.1 saying these descriptions come from vendor documentation and were not independently verified would cost one line and change how the section is read. Then strip the evaluative adjectives that came along with the copy ("robust suite," AWS entry, reference 7).

4. Major — an evaluative performance claim cited to a source that cannot contain it (Section 4.2.1, LangChain entry, reference 9).

"Faces challenges with speed limitations when managing complex interactions between multiple agents and tools" is attributed to the LangChain v0.1 agents documentation. Provide a measurement (version, task, hardware, concurrency, latency) or label it as the author's operational experience. Singling out one framework for a negative performance claim while describing the other five in their own promotional terms is also an even-handedness problem.

5. Major — novelty claim without prior art (Abstract; Section 3.1; Section 3.3.3).

Cognitive Skills is introduced as a new core building block and then defined in 3.3.3 as a "model hub" of purpose-built models the LLM dispatches to. The LLM-as-dispatcher-over-a-model-hub idea, and the tool-invocation formalisms it sits on, have an existing literature that is neither cited nor distinguished. Add a related-work paragraph and state the delta explicitly: what does Cognitive Skills specify that a typed tool registry with a router does not?

6. Major — no criterion separating Cognitive Skills from Tools (Sections 3.3.3 and 3.3.4; instantiated in Section 4.3.2).

RAG is a Tool, OCR is a Cognitive Skill, and both are remote calls returning a result. In the Figure 4 walkthrough, Search APIs (LLM Agent 3) are Tools while Guardrail Classifiers (LLM Agent 4) are Cognitive Skills, with no stated rule. Offer a decision rule — ownership of the model weights, whether the call performs inference versus retrieval, whether the artifact is trained on domain data — and then apply it consistently to every example already in the paper.

7. Major — the three categories are not disjoint (Section 4.3).

They are described as "three primary types based on their structure, scope, and interaction dynamics," but Human-Augmented is an oversight mode that can sit on top of either of the other two. The article's own Fraud Detection example (Section 4.3.3) is a multi-agent pipeline with a compliance officer in the loop. Either recut along one axis or present them as non-exclusive lenses and say which combinations occur.

8. **Major — no evaluation anywhere (whole paper).** No benchmark, no metric, no notion of what makes one pattern better than another for a given workload. A reader deciding between the Router pattern (Section 4.3.1) and the Orchestrated Multi-Agent pattern (Section 4.3.2) gets no cost, latency, error-propagation, or failure-mode comparison — and orchestration multiplies model calls, which is precisely the tradeoff a practitioner needs. Even a qualitative table with columns for call count, latency profile, failure modes, and when not to use it would be a substantial improvement.
9. **Moderate — "Practical Use Cases" contains no practice (Sections 4.3.1, 4.3.2, 4.3.3).** All fifteen or so entries are Scenario plus Solution, hypothetical throughout. Rename them ("Illustrative scenarios") or convert at least one per pattern into a real deployment with numbers. In the healthcare examples specifically (4.3.1 Healthcare Decision Support, 4.3.2 Healthcare Assistance, 4.3.3 Healthcare Diagnostics and Treatment Planning), an agent that "integrates results into a single, comprehensive report for the physician" (4.3.1) carries clinical and regulatory weight that the article does not acknowledge anywhere.
10. **Moderate — mis-citation (Section 3.2, reference 2).** "For more details on RAG implementation and advanced prompting guidelines, refer to [2] and [3]." Reference 3 (Gao et al., RAG survey, arXiv:2312.10997) is apt. Reference 2 is Giray, "Prompt engineering with ChatGPT: A guide for academic writers," *Annals of Biomedical Engineering* 51(12):2629–2633 — a guide to prompting for academic writing, not for agent or tool-use prompting. Replace it with prompting or tool-use work that matches the claim.
11. **Moderate — weak support for the opening premise (Section 1, reference 1).** The claim that "SaaS platforms becoming essential for scalability and efficiency across industries" is sourced to Cheng et al., an AIOps-on-cloud-platforms review (arXiv:2304.04661). That paper is about IT operations, not the SaaS market or its limits. The whole argument of Sections 1.1 and 1.2 — that SaaS and context-aware systems fall short in specific, named ways — is a claim about deployed software and has zero citations behind it.
12. **Moderate — the Anthropic agent-versus-workflow distinction is used before it is attributed (Sections 3.2 and 4.2.1).** Section 3.2 draws exactly the agents-versus-workflows line that Anthropic's "Building effective agents" (reference 8) is known for, but cites nothing; reference 8 appears later, in 4.2.1, for a different point. Move or duplicate the citation into 3.2. Reference 8 also lacks the access date that references 4 through 7 and 9 carry.
13. **Moderate — an uncited comparative empirical claim (Section 3.3.3).** "Purpose-built models specifically designed to accomplish tasks that general-purpose LLMs, even when fine-tuned LLMs,

struggle to perform effectively. Fine-tuned models often lack the precision and specialization required." This is the load-bearing justification for the paper's main contribution; it is contestable on several of the listed tasks (OCR and image classification in particular, given multimodal model progress over the review window), and it has no source and no measurement. It needs at least one citation, and ideally a task-by-task statement of where the gap does and does not hold.

14. **Minor — the article does not deliver the standardization it promises (Abstract versus Sections 4.2 and 4.3).** The Abstract says the work "attempts to define a level of standardization for Vertical AI agent design patterns"; Section 4.2 concedes no universal patterns exist; Section 5.1 then reports the goal as met, and Section 5.2 lists "developing standardized frameworks" as future work. Pick one position. If standardization is the contribution, give it a form another team can conform to — module interfaces, message schemas, a conformance checklist.
15. **Minor — figures and tables.** Figures 1 through 5 are referenced in Sections 3.1, 3.2, 4.3.1, 4.3.2, and 4.3.3; in the version I was given, only the captions came through, so I make no claims about their content. There are no tables at all. A framework comparison table (name, vendor or lab, version, access date, single-agent or multi-agent, orchestration model, license) would carry much of Section 4.2 far more usefully than the current bulleted prose.
16. **Minor — copy.** Section 2.1.3, final paragraph: "Vertical Ai agent solutions" — capitalization. The abstract uses "attempts to define," the conclusion uses the perfect tense for the same work; align them.

Recommendation

Major revisions — the article is a coherent and readable orientation to agentic design patterns, and I would be glad to see it in the literature, but in its current form, it makes outcome claims (Section 2, Section 5.1) with no citations, relays vendor self-description as finding (Section 4.2.1), and advances a new core module (Section 3.3.3) without prior-art comparison or a rule separating it from the Tools module; adding a stated review method, sourcing or downgrading the benefit claims, marking vendor material as vendor material, and positioning Cognitive Skills against existing model-routing and tool-use work would address the substance without requiring the paper to be rebuilt.

Declarations

Potential competing interests: No potential competing interests to declare.

