

Peer Review

Review of: "Physical AI Agents: Integrating Cognitive Intelligence with Real-World Action"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper proposes **Physical AI Agents** as a named category of embodied system that extends the author's earlier *Vertical AI Agents* paradigm (reference [6], the author's own arXiv preprint, which also supplies Figure 1) from purely digital, domain-specialised reasoning into direct interaction with the physical world. Two contributions are advanced. First, a modular reference architecture (Section 3.2, Figure 2) organised into three blocks — Perception (sensors: cameras, LIDAR, IMUs, IoT devices), Cognition (a fine-tuned domain LLM plus memory, cognitive skills and tools), and Actuation (robotic arms, grippers, hydraulics, mobility platforms) — offered explicitly to fill what Section 3.2 calls the absence of “a universally accepted architecture.” Second, a design pattern named **Ph-RAG** (Physical Retrieval-Augmented Generation, Section 4.2, Figure 3), in which an embodied agent operating in the “real-world ecosystem” supplies processed physical context to an external industry-specific LLM operating in the “knowledge ecosystem,” which then performs higher-order reasoning and produces reports. The claims are supported by five short application vignettes (Section 3.3: autonomous vehicles, warehouse robotics, healthcare robotics, manufacturing, agriculture), a survey of vendor platforms (Section 3.4), and two extended case studies — oil and gas pipeline integrity monitoring (Section 4) and a hybrid Vertical-plus-Physical warehouse replenishment system (Section 5). The abstract states that these case studies “demonstrate how Physical AI Agents and the Ph-RAG framework are transforming industries.”

2. Whether the conclusion holds on the evidence

Nothing in the paper is measured. The article contains three figures (Figures 1–3), no tables, and no numerical results of any kind. The only two quantities anywhere in the text are scenario stipulations, not findings: “over 50,000 products” in Section 5.1 and a forecasting horizon of “30 minutes” in Sections 5.3 and 5.4. Neither case study reports a deployment, a simulation, a benchmark, a baseline, a latency budget, an accuracy figure, or a date. Section 4 is organised as “Problem Statement → Solution Design → Components → End-to-End Workflow,” and Section 5 likewise — these are design narratives, and the “Example Flow” bullets in Section 3.3 are four-line stories (Input → Perception → Reasoning → Action). A conceptual paper is entitled to argue rather than measure, but it is not entitled to the verbs “demonstrate” and “are transforming” in the abstract, or “the Ph-RAG framework is versatile and scalable” in Section 4.5, on this basis. Section 4.5 in particular generalises to infrastructure inspection, environmental monitoring, and precision agriculture from a single worked-out hypothetical.

The load-bearing sections are citation-free. Of 15 reference entries, five ([1]–[5]) appear in Section 1.1, two occurrences of [6] appear in Section 2.1 and the Figure 1 caption, and nine ([7]–[15]) appear in Section 3.4. Sections 3.1, 3.2, 3.3, 4.1–4.5, 5.1–5.4, and 6.1–6.2 contain **zero citations**. That is the entire proposed contribution: the definition, the three-block architecture, Figure 2, all five application vignettes, both case studies, Figure 3, and the future-directions agenda. The most consequential instance is the gap claim in Section 3.2 — “there is no universally accepted architecture for the design and implementation of Physical AI Agents” — which is the justification for the paper’s central contribution and is asserted without a single supporting reference. A gap claim is exactly the sentence that a conceptual paper must evidence with a survey, because the architecture proposed (sense → plan → act) is the oldest decomposition in autonomous robotics, and the paper nowhere states what distinguishes its version from that tradition.

Vendor material is doing the work of evidence. Section 3.4, the paper’s only empirical survey of the field, rests on nine references, none of which is a research publication: [7] NVIDIA Isaac product page, [8] ros.org, [9] AWS RoboMaker product page, [10] robotics.google.com, [11] microsoft.com/robotics, [12] the CES 2025 event website, [13] a *Wired* news story standing in for the technical description of NVIDIA Cosmos, and [14] and [15] — which are **the same Business Insider stock-market news URL cited twice under two different labels** (Edify SimReady and Omniverse Blueprints), so the 15 entries resolve to 14 distinct sources. The substantive technical claim in Section 3.4 — that Cosmos enhances robot understanding of the physical world “through synthetic data generation” — is sourced to consumer

press, not to a model card or technical report. Across the whole list, exactly one entry ([1], Goertzel & Wang, IOS Press, 2007) is a formally published peer-reviewed venue; five ([2]–[6]) are arXiv preprints, one of them the author’s own; the remaining nine are vendor pages and trade press. **No robotics, embodied-AI, vision-language-action, or robot-learning research literature is cited at all.** For a paper whose thesis is that LLM cognition should be placed inside a perception-action loop, this is the decisive omission: the reader cannot tell whether the proposal restates a solved problem, contradicts known results, or is genuinely new.

The taxonomy has no discrimination rule, and the paper’s own material shows it. Section 3.1 defines a Physical AI Agent as an embodied system with sensory capability, cognitive intelligence, and actuation, and adds that it is “equipped with industry-specific intelligence, powered by fine-tuned, domain-specific LLMs.” Nowhere does the paper state a test that lets a reader classify a borderline system. Three borderline cases arise from the paper’s own text:

(a) *Section 3.3, Autonomous Vehicles.* The flagship example is a self-driving car with LIDAR, cameras, radar, GPS, path planning, collision avoidance, and braking actuators. A conventional autonomous-vehicle stack has all of these and no LLM. By the Section 3.1 definition, which makes a fine-tuned domain LLM constitutive, such a vehicle is *not* a Physical AI Agent — yet Section 3.4 presents ROS [8], Isaac [7], and RoboMaker [9], all pre-LLM platforms, as the foundational tooling for exactly these agents. The definition in Section 3.1 and the examples in Sections 3.3 and 3.4 cannot both be right, and the reader is given no rule to resolve them.

(b) *Section 5.3, the two component lists side by side.* The Vertical AI Agent and the Physical AI Agent are each specified with Cognitive Skills, Tools, a Reasoning Module (“specialized LLM”), and a Memory Module; the Physical agent adds Sensors and Actuators. The sole discriminator is therefore possession of actuators. But Section 4.3 classifies aerial drones and ground robots as “Actuators,” i.e., locomotion counts as actuation — in which case the supply-chain Vertical AI Agent of Section 2.3, which “dynamically adapts to disruptions” by issuing instructions that move trucks and stock, actuates the physical world too. The paper never says whether the boundary is contact-level manipulation, closed-loop physical control, co-location with the effector, or something else.

(c) *Section 3.3, Agriculture.* The actuators listed are “autonomous harvesters, irrigation systems, and pesticide sprayers.” A fixed irrigation valve opened on an LLM’s recommendation satisfies every clause of the Section 3.2 architecture — sensors (soil moisture), cognition (fine-tuned agricultural LLM), actuation

(valve). Under this reading, an LLM-controlled thermostat is a Physical AI Agent. If that is intended, the category is too broad to do work; if it is not, Section 3.2 must state the excluding condition.

Ph-RAG is not shown to involve retrieval. In retrieval-augmented generation, a query is embedded, an index is searched, and the returned passages condition generation. In Sections 4.2 and 4.4, the flow is the opposite: the Physical AI Agent *pushes* processed sensor context to the external LLM, which then reports. Figure 3, the only specification of the pattern, defines no index over physical observations, no retriever, no query representation, no similarity criterion, and no relevance or freshness policy. The “vector search” that does appear (Sections 3.2, 4.3, and 5.3) is conventional document retrieval inside the Tools module and is orthogonal to the physical channel. As presented, Ph-RAG is a sensor-to-LLM context pipeline, and the name asserts a mechanism the paper does not define. This matters beyond terminology: the interesting questions a physical RAG would raise — what is indexed, at what temporal granularity, how stale physical observations are invalidated, how contradictory sensor readings are reconciled — are precisely the ones the pattern would need to answer to be a contribution rather than a label.

Safety-critical claims are made without qualification. Section 3.3 (Healthcare Robotics, Example Flow) places the LLM in the loop that determines “the optimal incision path,” and Section 3.3 (Autonomous Vehicles, Example Flow) places it in the loop that brakes for a crossing pedestrian. Neither passage discusses latency, determinism, hallucination, verification, fallback control, or regulatory approval. Section 6.2 later acknowledges that LLM quantisation and latency reduction for resource-constrained robotic platforms remain open problems — which concedes that the loops asserted in Section 3.3 are not currently realisable as described. A conceptual paper may propose such architectures, but it must mark them as proposals and discuss their failure modes; here they are narrated as working systems.

3. Reproducibility: the review protocol

For a paper of this type, reproducibility means that an independent reader could repeat the literature survey and arrive at the same landscape, and could apply the taxonomy to new systems and arrive at the same classifications. Neither is possible here.

No review method is stated anywhere. Section 1.3 (“Purpose of This Work”) lists three objectives but specifies no search base (Scopus, Web of Science, IEEE Xplore, arXiv, Google Scholar), no time window, no query strings, no inclusion or exclusion criteria, and no screening or selection procedure. The implicit cutoff is January 2025 — inferable only from [12] and [13] referring to CES 2025 and from the submission date — but this is never declared. Section 3.4 asserts that platform development “is in its early stages,

with several industry leaders making notable strides” and then names five platforms ([7]–[11]); no rule is given for why these five and not others, so a second reader cannot reproduce the selection. The claim that the field is “in its early stages” is itself a survey finding presented without a survey.

Reference count and clustering. Fifteen numbered entries, 14 distinct sources ([14] and [15] share one URL). Distribution by section: Section 1.1 → [1]–[5] (5 entries); Section 2.1 and the Figure 1 caption → [6] twice (1 entry, marked “a, b”); Section 3.4 → [7]–[15] (9 entries); every other section → none. Thus **14 of the 15 entries sit in two sections (1.1 and 3.4) that contain none of the paper’s contribution**, while Sections 3.1–3.3, 4, 5, and 6 — which contain all of it — cite nothing. By source type: 1 peer-reviewed book ([1], 2007), 5 arXiv preprints ([2]–[6], including the author’s own [6]), 5 vendor product pages ([7]–[11]), 1 trade-event website ([12]), and 3 press items ([13] *Wired*, [14], and [15] *Business Insider*). Fifteen references are thin for a framework paper that claims to identify a gap in the state of the art, and the composition — 60% vendor and press — means the paper’s picture of the field is assembled from marketing material.

The taxonomy is not reproducibly applicable. As shown in Section 2 above, the classification of the autonomous vehicle (Section 3.3), the Vertical/Physical boundary (Section 5.3 against Section 4.3), and the irrigation controller (Section 3.3) cannot be settled from the text. Two readers applying Section 3.2 to the same system may reach opposite verdicts, which is the operational definition of a taxonomy that has not yet been specified.

Figures. Figure 1 is reproduced from [6]; Figures 2 and 3 are block diagrams whose captions carry no legend explaining arrow semantics (data flow? control flow? invocation?), no interface definitions between blocks, and no indication of which arrows are synchronous. A reader cannot implement from them.

4. What specifically to fix

1. **Add a review protocol subsection to Section 1.3.** State the databases searched, the exact query strings, the date window, the cutoff date, the inclusion and exclusion criteria, the number of records screened and retained, and how [7]–[11] were selected for Section 3.4. Without this, Section 3.4’s “early stages” finding is unverifiable.
2. **Substantiate the gap claim in Section 3.2.** The sentence “there is no universally accepted architecture for the design and implementation of Physical AI Agents” is the paper’s justification for existing and currently carries no reference. Support it with a survey of the existing architectural proposals it displaces, and state explicitly how the Perception/Cognition/Actuation decomposition

of Figure 2 differs from the classical sense–plan–act pipeline and from subsumption and hybrid deliberative–reactive architectures.

3. **Cite the embodied-AI and robot-learning literature.** The reference list contains no research publication in robotics, manipulation, embodied agents, or vision-language-action models. Sections 3.1, 3.2, and 4.2 cannot be assessed as novel until it is shown what existing work on LLM-driven robot control they build on or contradict.
4. **Replace press and product pages with primary technical sources in Section 3.4.** Specifically: [13] cites *Wired* for the technical capability of NVIDIA Cosmos; [14] and [15] cite the same Business Insider stock-news URL for two distinct products (Edify SimReady and Omniverse Blueprints). Cite model cards, technical reports, or documentation, and correct the duplicate so that 15 entries do not collapse to 14 sources.
5. **State a discrimination rule for the taxonomy in Section 3.2, and apply it to at least three borderline cases.** Use the paper’s own material: (i) the conventional non-LLM autonomous vehicle implied by Section 3.3 and by ROS/Isaac/RoboMaker in Section 3.4 — is it a Physical AI Agent, given that Section 3.1 makes a fine-tuned LLM constitutive? (ii) the fixed irrigation valve of Section 3.3 (Agriculture), which satisfies every clause of Figure 2; (iii) the supply-chain agent of Section 2.3, which changes physical states through orders. Say whether the criterion is contact-level manipulation, closed-loop physical control, or embodiment, and make it decide each case.
6. **Make the Vertical/Physical boundary explicit in Section 5.3.** The two component lists differ only by Sensors and Actuators, while Section 4.3 counts drones and ground robots as Actuators. Add one sentence stating the necessary and sufficient condition, and reconcile it with Section 4.3.
7. **Specify the retrieval mechanism of Ph-RAG in Section 4.2 and Figure 3, or rename the pattern.** If it is retrieval, define what is indexed (raw observations? anomaly records? spatial-temporal keys?), the query representation, the similarity criterion, the freshness and invalidation policy for stale physical observations, and the conflict-resolution rule for contradictory sensor readings. If it is a push pipeline from agent to LLM, as Sections 4.2 and 4.4 describe, say so and choose a name that does not assert retrieval — and distinguish it explicitly from the conventional “vector search” already listed in the Tools modules of Sections 3.2, 4.3, and 5.3.
8. **Align the abstract and Section 4.5 with what the paper actually shows.** Replace “we demonstrate how Physical AI Agents and the Ph-RAG framework are transforming industries” (abstract) and “the Ph-RAG framework is versatile and scalable” (Section 4.5) with language that identifies Sections 4

and 5 as illustrative designs. Either qualify the claims or supply at least one instantiation with measurements.

9. **Add at least one evaluable instantiation, however small.** A single simulated pipeline-inspection run in an existing simulator, reporting end-to-end latency from anomaly detection (Section 4.4, step 2) to report delivery (step 5), plus detection precision against a labelled set, would convert Section 4 from narrative to evidence. Alternatively, state falsifiable predictions that the framework makes and that a competing architecture does not.
10. **Qualify the safety-critical loops in Section 3.3.** For Healthcare Robotics, the Example Flow has the LLM determine “the optimal incision path”; for Autonomous Vehicles, it has the LLM in the braking decision for a pedestrian. Add a subsection on failure modes, latency budgets, determinism requirements, hallucination mitigation, human oversight, and certification, and reconcile these with the acknowledgement in Section 6.2 that on-device LLM latency is still an open problem.
11. **Define the interfaces between the three blocks of Figure 2.** Give the arrows a legend (data versus control versus invocation), state what representation crosses Perception → Cognition and Cognition → Actuation, and indicate which paths are synchronous. Do the same for the two ecosystems in Figure 3. As drawn, neither figure is implementable.
12. **Separate the contribution from prior work in Section 2.** Section 2 restates [6], and Figure 1 is taken from it. State at the head of Section 3 exactly which elements are new here as against [6], so a reader can see the increment.
13. **Reconcile the Disclaimer with the stated CC BY licence.** The Statements and Declarations assert that the architectures, diagrams, and concepts “are the intellectual property of the author” and that attribution “is expected,” which reads as more restrictive than the CC BY terms under which the article is published. Clarify.

5. Recommendation

Major revisions. The paper is clearly written, well organised, and addresses a question that deserves a framework paper: what changes when domain-specialised LLM cognition is placed inside a perception-action loop. The three-block architecture of Figure 2 and the Ph-RAG pattern of Figure 3 are legitimate conceptual objects. But in its present form, the work does not meet the standards that apply to conceptual papers specifically. It states no review method (Section 1.3); it grounds its central gap claim in Section 3.2 with no citation; it builds its picture of the field in Section 3.4 out of five vendor pages, an event site, and three press items, one of them duplicated ([14] and [15]); it cites no robotics or embodied-

AI research at all, so novelty cannot be assessed; its taxonomy provides no rule that decides the borderline cases its own Sections 3.3, 4.3, and 5.3 generate; and its central pattern is named for a retrieval mechanism that Figure 3 does not define. The abstract's "demonstrate" overstates two hypothetical case studies whose only numbers are a stipulated 50,000 SKUs and a stipulated 30-minute horizon (Sections 5.1, 5.3, 5.4). Items 1, 2, 3, 5, 7, and 8 above are, in my view, necessary before the work can support its conclusions; the remainder would substantially strengthen it. None of this requires the paper to become an empirical study — it requires it to be a rigorous conceptual one.

Declarations

Potential competing interests: No potential competing interests to declare.