CEA Comité Español de Automática
XLVII JORNADAS DE AUTOMÁTICA CÓRDOBA 2026
UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Estructura antes que escala: hacia la integración fiable de agentes de lenguaje en la industria

Gómez-González, P.-P.[a,b,*], Freire-Mahía, N.[a], Michelena, Á.[a], Rodríguez-Olmo, D.[b], Jove, E.[a], Calvo-Rolle, J. L.[a]

[a] *Universidade da Coruña, Research Group CTC, CITIC, Departamento de Enxeñaría Industrial, Campus de Ferrol, 15403 Ferrol, A Coruña, Spain.*
[b] *Técnicas de Soft, S.A., Polígono de Pocomaco, 5ª avenida, parcela E-18, 15190 A Coruña, Spain.*

To cite this article: Gómez-González, P.-P., Freire-Mahía, N., Michelena, Á., Rodríguez-Olmo, D., Jove, E., Calvo-Rolle, J. L. 2026. Structure before scale: toward reliable integration of language agents in industry. Jornadas de Automática, 47. https://doi.org/10.17979/ja-cea.2026.47.13835

Resumen

Los agentes basados en modelos de lenguaje se proponen para operar e interrogar sistemas industriales, pero su adopción tropieza con un obstáculo persistente: la escala del modelo mejora la fiabilidad, pero no hasta el nivel que un entorno industrial exige, donde una decisión errónea puede acarrear pérdidas. El presente trabajo sostiene una posición, estructura antes que escala, según la cual, sin restar valor a la capacidad del modelo, la mayor palanca de fiabilidad en planta es la estructura impuesta sobre él, que además no exige el coste de infraestructura de un modelo mayor. La posición se articula a partir de literatura reciente, que converge en tres habilitadores (grafos de conocimiento, agentes y el protocolo MCP) y en un techo de fiabilidad conocido, y se ilustra con un despliegue propio: la operación de un sistema SCADA mediante un agente y MCP, donde encapsular la coordinación frágil en código reduce las invocaciones del modelo de entre 20 y 50 a una sola.

Palabras clave: Sistemas multiagente, Soporte al operador humano, Protocolos de comunicación industrial, Sistemas de control en red seguros, Inteligencia artificial

Structure before scale: toward reliable integration of language agents in industry

Abstract

Language-model agents are increasingly proposed to operate and query industrial systems, yet their adoption meets a persistent obstacle: scaling improves reliability, but not to the level an industrial environment demands, where a wrong decision can cause losses. This work argues a position, structure before scale, whereby, without denying the role of model capability, the greatest lever for reliability on the plant floor is the structure imposed on the model, which moreover does not require the infrastructure cost of a larger model. The position is built from recent literature, which converges on three enablers (knowledge graphs, agents and the MCP protocol) and on a well-documented reliability ceiling, and is illustrated with one deployment: operating a SCADA system through an agent and MCP, where encapsulating brittle coordination in code cuts model invocations from 20–50 to a single one.

Keywords: Multi-agent systems, Human operator support, Industrial communication protocols, Secure networked control systems, Artificial intelligence

1. Introducción

Los modelos de lenguaje de gran tamaño (LLM) ofrecen, a la vez, una promesa y una advertencia. Por un lado, resuelven con soltura tareas que la automatización clásica abordaba con dificultad: resumen de documentación técnica, traducción, generación de código o interacción en lenguaje natural con sistemas complejos. Por otro, arrastran una lista bien documentada

*Autor para correspondencia: pedro.pablo.gomez@udc.es

de comportamientos poco fiables, desde las alucinaciones hasta la imposibilidad de explicar sus propias decisiones, que ha llevado a cuestionar si satisfacen los estándares de ingeniería exigibles a un sistema automatizado de confianza (Denning, 2026). Esa tensión entre lo que prometen y lo que garantizan es el punto de partida de su llegada a la industria.

La pregunta relevante, por tanto, no es si un agente basado en LLM puede operar o interrogar un sistema industrial, pues la respuesta empírica es afirmativa, sino cómo debe hacerlo para ser fiable. El presente trabajo sostiene una posición al respecto, que puede resumirse como estructura antes que escala: en un entorno de tecnología operacional (OT), la palanca de mayor rendimiento para la fiabilidad no es la capacidad del modelo, sino la estructura que se impone sobre él. No se niega que un modelo mayor sea más capaz y, en alguna medida, más fiable; lo que se constata es que esa mejora no alcanza el nivel que la industria exige, donde una decisión errónea puede traducirse en pérdidas materiales, de proceso o de seguridad. Frente a ello, acotar el comportamiento del modelo mediante código, esquemas, herramientas de alto nivel y protocolos bien diseñados aporta fiabilidad de forma incremental y, sobre todo, sin el coste de infraestructura que implica desplegar un modelo mayor.

La posición se articula en tres pasos. La Sección 2 revisa el estado del arte y su trayectoria, de la orquestación artesanal a unos habilitadores ya estandarizados, y el techo de fiabilidad que la escala no levanta. La Sección 3 la ilustra con un despliegue propio sobre un sistema SCADA (*Supervisory Control and Data Acquisition*) real. La Sección 4 proyecta hacia dónde se dirige el campo, leyendo la literatura reciente en clave prospectiva. Conviene precisar el alcance de la aportación: este es un artículo de posición cuya contribución principal es conceptual y arquitectónica, ilustrada mediante un despliegue real desarrollado por los autores, y no una validación experimental exhaustiva ni una revisión sistemática de la literatura.

2. Estado del arte y su trayectoria

2.1. De la orquestación artesanal a los habilitadores estándar

Los primeros asistentes industriales basados en LLM resolvían la integración de forma artesanal. Un ejemplo representativo combinaba la recuperación sobre documentación de máquina con un LLM de capacidad comparable a GPT-3.5 y un orquestador construido a mano para delegar en herramientas externas (Garcia et al., 2024). Ya entonces se observaba que el modelo, por sí solo, fallaba en tareas tan básicas como sumar de forma consistente, lo que obligaba a externalizar el cálculo. Dos años después, la literatura ha sistematizado ese patrón. Una revisión reciente identifica tres habilitadores para el despliegue de LLM en fabricación inteligente: los grafos de conocimiento, los sistemas de agentes y el protocolo MCP (Model Context Protocol) (Chen et al., 2026). Este último estandariza la comunicación entre el modelo y las herramientas externas y, al hacerlo, aborda problemas de fragmentación de datos, latencia e interpretabilidad, ofreciendo un lenguaje común entre humanos, máquinas y software que, de otro modo, dependen de herramientas e interfaces dispares y mal comunicadas entre sí. Lo que antes se resolvía a mano es hoy una pieza con nombre propio.

2.2. El techo de fiabilidad

Esa madurez de las herramientas convive con un techo de fiabilidad que la escala mejora, pero no llega a levantar. Un estudio a gran escala sobre múltiples familias de modelos matiza la intuición habitual: aunque los modelos mayores y más instruidos aciertan más, no aseguran regiones de operación libres de error, ni siquiera en las tareas sencillas. A ello se añade un efecto que los vuelve, en cierto sentido, menos fiables: entrenados para no eludir preguntas, dejan de abstenerse cuando no saben y, en lugar de reconocer su desconocimiento, devuelven respuestas plausibles pero incorrectas, un tipo de error más insidioso que, además, la supervisión humana no consigue compensar (Zhou et al., 2024). La valoración desde la ingeniería es igual de severa: los LLM presentan especificaciones difusas, no permiten saber con certeza si funcionan correctamente y no explican cómo llegan a sus respuestas, de modo que no se les están aplicando los métodos estándar para asegurar la confianza en sistemas automatizados (Denning, 2026). En los agentes que invocan herramientas, este techo adopta una forma caracterizada: las alucinaciones en la ejecución de herramientas, que la literatura distingue entre errores en la selección de la herramienta y errores en su invocación, cuando el modelo genera argumentos sintácticamente válidos pero referidos a entidades inexistentes (Peng et al., 2026).

2.3. El hueco de seguridad

A la fiabilidad se suma la seguridad. La apertura que convierte a MCP en un estándar útil es, al mismo tiempo, una superficie de ataque. La literatura ha respondido con soluciones de seguridad, como un *middleware* que añade autenticación, limitación de tasa y filtrado de peticiones (Kumar et al., 2025), o marcos de mitigación de grado empresarial basados en principios de confianza cero (Narajala and Habler, 2025). Son aportaciones valiosas, pero comparten un rasgo que deja al descubierto dos carencias: actúan sobre la infraestructura de TI que rodea al protocolo, no sobre el proceso industrial. De ahí la primera carencia, de enfoque, pues el riesgo crítico de MCP en la industria son las acciones físicas sobre los sistemas conectados (Narajala and Habler, 2025), y ese riesgo no se captura tratando una orden insegura como un patrón de texto sospechoso. La segunda carencia es empírica: el marco más completo admite que su validación es conceptual y que falta evidencia de campo sobre la efectividad real de las contramedidas (Narajala and Habler, 2025), en un terreno donde se ha llegado a cifrar en torno al 43 % los servidores MCP vulnerables a inyección de comandos (Kumar et al., 2025).

En suma, el campo conoce sus habilitadores y reconoce su techo, pero carece todavía de un relato estructural de cómo obtener fiabilidad en planta y de evidencia que lo respalde. Esa es la brecha que motiva la posición de este trabajo.

3. Caso de estudio y resultados: agente, MCP y SCADA

Para aterrizar la posición conviene un caso concreto. El despliegue y los resultados que se resumen a continuación son fruto del presente trabajo, desarrollado por los autores, y no un ejemplo tomado de la literatura: un agente basado en LLM integrado con un sistema SCADA real, una instalación eléctrica de media tensión con una veintena de analizadores de red en celdas de 6 y 15 kV.

Dos enfoques compiten para esta tarea. El primero son las *skills*: instrucciones en lenguaje natural que orquestan herramientas genéricas de bajo nivel y delegan en el modelo la coordinación entre llamadas. El segundo expone herramientas de alto nivel específicas del dominio, en las que una sola invocación produce el resultado completo y toda la coordinación reside en código. Para materializar este segundo enfoque, se ha implementado una arquitectura estructurada en capas, ilustrada en la Figura 1. En lugar de exponer directamente el servidor genérico del SCADA al modelo, se introduce un Servidor MCP Propio que actúa como intermediario.

Como se observa en la topología (Figura 1), este servidor intermedio absorbe la complejidad operativa aislando al cliente LLM de los datos crudos. La distinción importa para no atribuir a MCP un mérito que no le corresponde: la ganancia no proviene del protocolo en sí, sino de retirar del modelo la manipulación de la lógica frágil. Tareas críticas como la paginación de respuestas extensas, el manejo de errores del sistema de control y el formateo masivo de datapoints se resuelven de forma determinista en el código del intermediario. MCP es únicamente la interfaz estándar que hace esa solución limpia y reutilizable.

El porqué de esta estructura se entiende al observar el modo de fallo clásico del primer enfoque. Las rutas canónicas de las variables SCADA son cadenas largas y repetitivas, y el modelo las corrompe al transcribirlas entre llamadas (por ejemplo, `833` transcrito como `8MS`, o `DI3` como `DIS`), produciendo consultas inválidas que no generan señales evidentes de error. Es una instancia exacta de las alucinaciones en la invocación de herramientas (Peng et al., 2026) y de la recuperación probabilística que describe la literatura (Denning, 2026). La solución, por tanto, es estructural: retirar la manipulación de estas cadenas del dominio probabilístico del modelo y encapsularla en la capa de herramientas de alto nivel del servidor intermediario.

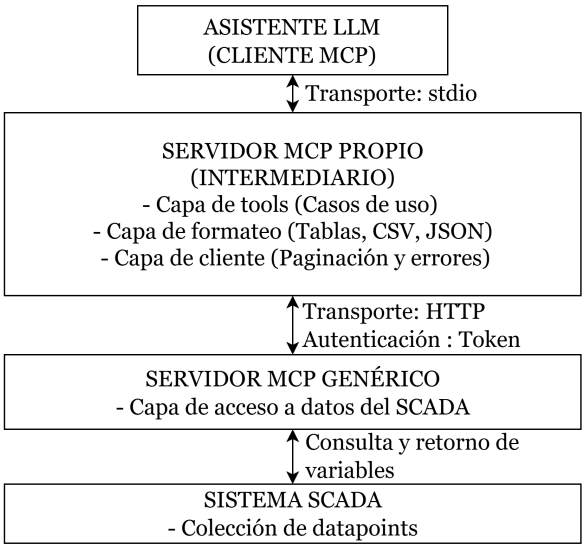

Figura 1: Topología de integración del agente LLM con el sistema SCADA mediante un servidor MCP intermediario que expone herramientas de alto nivel.

El efecto se resume en la Tabla 1. El modelo, un Gemma 4 26B, se ejecutó de forma local en una estación de ingeniería sin recursos especializados de cómputo (Intel Core Ultra 7 265, 64 GB de RAM, NVIDIA RTX 4000 Ada). Los valores de la tabla no proceden de una batería de pruebas controlada, sino de la observación del sistema durante su desarrollo y uso sobre la instalación real. Las consultas, representativas de la operación diaria, fueron de tres tipos: enumeración de los analizadores de la instalación, lectura de las magnitudes eléctricas de una celda concreta y localización del *datapoint* asociado a una variable de planta. Las invocaciones y los tiempos de respuesta se expresan como rangos observados en ejecuciones repetidas de esas consultas con ambos enfoques, mientras que la corrupción de identificadores y la detección de fallos parciales se registraron cualitativamente como presencia o ausencia. El caso no demuestra una ley general, pues se limita a una única instalación (una red de distribución de media tensión con una veintena de analizadores) y a una única familia de tareas (consultas de lectura sobre datos de proceso en tiempo real, sin acciones de escritura sobre la planta), pero ilustra con nitidez el mecanismo que la posición postula.

Tabla 1: Orquestación por *skills* frente a herramientas de alto nivel sobre MCP en el caso de estudio.

Aspecto	*Skills*	MCP propio
Invocaciones del LLM / consulta	20–50	1
Tiempo de respuesta	∼25 min	4–6 min
Corrupción de identificadores	recurrente	ausente
Detección de fallos parciales	no	sí
Reproducibilidad	baja	alta

Conviene reconocer el coste de este enfoque. Trasladar responsabilidades del modelo al código añade trabajo de ingeniería y resta parte de la versatilidad inmediata que hace atractivos a los LLM, pues obliga a estructurar con detalle tareas que el modelo abordaría de forma genérica y puede exigir rehacer esa estructura cuando el proceso con el que se interactúa cambia. Es una contrapartida real. Pero en un entorno industrial la fiabilidad no es negociable, y ese coste se asume como parte del precio de operar con garantías. Buena parte de la evolución reciente del campo, de hecho, puede leerse como el esfuerzo por abaratar esa estructura sin renunciar a la fiabilidad.

4. Tendencias y perspectivas

Si la fiabilidad se apoya sobre todo en la estructura, la pregunta natural es hacia dónde conduce esa idea. La literatura reciente, leída en clave prospectiva, dibuja cuatro direcciones que comparten un mismo movimiento: estructurar y distribuir antes que confiarlo todo a la escala.

4.1. Modelos locales en el borde

La primera es el desplazamiento del cómputo hacia el borde. La restricción de las leyes de escalado, donde desplegar un modelo mayor incrementa el coste organizativo sin garantía de fiabilidad (Chen et al., 2026), empuja hacia LLM ligeros optimizados para el *edge* mediante mezcla de expertos, destilación o cuantización, que reducen la latencia y mantienen los datos sensibles fuera de servidores externos (Chen et al., 2026). El argumento de privacidad es explícito: una nueva generación

de modelos locales que operan sin conexión de red protege frente a la fuga de datos a la nube, aunque sin toda la potencia de un LLM (Denning, 2026). El despliegue descrito, con un modelo local en una estación de ingeniería, es coherente con esta dirección y sugiere que el compromiso entre capacidad y control vuelve a resolverse con estructura.

4.2. Colaboración entre modelos

La segunda es la colaboración entre modelos. En lugar de un único modelo que lo resuelva todo, la tendencia apunta a combinar un modelo grande de razonamiento general con modelos pequeños y especializados, por ejemplo redes neuronales informadas por física para el cálculo numérico preciso, coordinados mediante un protocolo como MCP (Chen et al., 2026). MCP deja así de ser solo una pasarela hacia el sistema heredado para convertirse en el tejido de coordinación entre componentes especializados, y la estructura pasa a estar distribuida entre ellos. Es una versión más ambiciosa del mismo principio que el caso de estudio ilustra a pequeña escala.

4.3. Seguridad como semántica de proceso

La tercera dirección es de seguridad, y el hueco que la literatura reconoce no se cierra trasladando a OT las defensas de TI. La acción insegura relevante en un proceso industrial es semántica: una orden puede ser sintácticamente válida y, aun así, físicamente peligrosa, de modo que la seguridad debe razonar sobre el significado de la acción para el proceso controlado y no sobre patrones de texto. Avanzar exige, como señala la propia literatura, métricas de seguridad medibles para MCP en entornos OT (Narajala and Habler, 2025), así como defensas frente a amenazas que hoy siguen abiertas, como la inyección de instrucciones a través del contenido de la propia fuente de datos (Kumar et al., 2025; Denning, 2026).

4.4. Hacia el diagnóstico explicable y causal

La cuarta dirección devuelve el foco a la confianza del operario, pues un agente que actúa sobre planta no solo debe acertar, sino justificar por qué actúa. Aquí la literatura apunta a la inferencia causal: dotar a los modelos de estructuras causales explícitas para pasar de la correlación a la causa en el diagnóstico de fallos (Chen et al., 2026). Combinar la atribución explicable, que cuantifica qué variables contribuyen a una anomalía, con modelos causales que distingan la causa del mero síntoma es una línea de trabajo natural para los agentes industriales, y un terreno en el que la estructura vuelve a preceder a la capacidad. Converge, además, con propuestas que ya hoy intercalan recuperación y verificación para acotar al modelo en dominios de decisión complejos (Zhou et al., 2025).

5. Conclusiones

Este trabajo ha defendido una posición, estructura antes que escala, sobre la integración de agentes de lenguaje en la industria. La literatura reciente converge en los habilitadores y reconoce el techo de fiabilidad; un despliegue sobre un sistema SCADA real ilustra que la fiabilidad se obtiene retirando la lógica frágil del modelo y encapsulándola en código expuesto mediante MCP; y la trayectoria del campo, leída en su conjunto, apunta de forma consistente a estructurar y distribuir, mediante modelos locales, colaboración entre modelos, seguridad de proceso y diagnóstico causal, antes que a depender de la mera escala. Para la comunidad de Automática, la implicación es que la integración fiable de la IA generativa en OT es, ante todo, un problema de ingeniería de la estructura que rodea al modelo, más próximo al diseño de la supervisión y de la tolerancia a fallos que a la elección del modelo de mayor capacidad. Como trabajo futuro, se extenderá la validación a otras instalaciones y a familias de tareas adicionales, se cuantificará de forma sistemática la fiabilidad de ambos enfoques mediante baterías de consultas repetidas con criterios de éxito predefinidos, y se abordarán las direcciones esbozadas en la Sección 4, en particular la seguridad entendida como semántica de proceso y la incorporación de mecanismos de verificación previos a cualquier acción de escritura sobre la planta.

Agradecimientos

La investigación de Pedro-Pablo Gómez-González ha sido financiada por la Xunta de Galicia mediante el programa Doutoramento Industrial (`http://gain.xunta.gal`), en el marco de la convocatoria Doutoramento Industrial 2025 con referencia 21_*IN*606*D*_2025_2261362.

El CITIC, como centro con acreditación en excelencia del Sistema universitario de Galicia y miembro de la Red CIGUS, es subvencionado por la Consellería de Educación, Ciencia, Universidades y Formación Profesional de la Xunta de Galicia y cofinanciado a través del Fondo Europeo de Desarrollo Regional (FEDER), Programa operativo FEDER Galicia 2021-27, (Ref. *ED*431*G* 2023_01)

Proyecto PID2022-137152NB-I00 financiado por MICIU/AEI/10.13039/501100011033 y por ERDF/EU.

This research is co-financed by the Interreg Atlantic Area Programme through the European Regional Development Fund, *EAPA*_0019_2022 SAtComm project.

Referencias

Chen, C., Li, L., Wang, T., Wang, Z., Leng, J., Shao, H., Cheng, L., 2026. Leveraging large language models for smart manufacturing: Reviews, enablers, challenges, and opportunities. Journal of Manufacturing Systems 85, 593–612.
DOI: 10.1016/j.jmsy.2026.02.008

Denning, P. J., 2026. The conundrum of LLMs. Communications of the ACM 69 (3), 31–34.
DOI: 10.1145/3789199

Garcia, C. I., DiBattista, M. A., Letelier, T. A., Halloran, H. D., Camelio, J. A., 2024. Framework for LLM applications in manufacturing. In: Manufacturing Letters (52nd SME North American Manufacturing Research Conference, NAMRC 52). Vol. 41. pp. 253–263.
DOI: 10.1016/j.mfglet.2024.09.030

Kumar, S., Girdhar, A., Patil, R., Tripathi, D., 2025. MCP Guardian: A security-first layer for safeguarding MCP-based AI systems. ArXiv:2504.12757.
DOI: 10.48550/arXiv.2504.12757

Narajala, V. S., Habler, I., 2025. Enterprise-grade security for the Model Context Protocol (MCP): Frameworks and mitigation strategies. ArXiv:2504.08623.
DOI: 10.48550/arXiv.2504.08623

Peng, H., Zheng, Y., Liu, Z., Lam, K.-Y., 2026. Tool execution hallucination in llm-based agents: A unified taxonomy with detection, mitigation, and future directions. TechRxiv 2026 (0227).

URL: https://www.techrxiv.org/doi/abs/10.36227/techrxiv.177219979.94060974/v1
DOI: 10.36227/techrxiv.177219979.94060974/v1

Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., Hernández-Orallo, J., 2024. Larger and more instructable language models become less reliable. Nature 634 (8032), 61–68.
DOI: 10.1038/s41586-024-07930-y

Zhou, Q., Li, M., Li, W., Qu, T., 2025. LLM-assisted advanced planning and scheduling framework for smart manufacturing systems. In: Proc. 17th International Conference on Computer Modeling and Simulation (ICCMS). pp. 162–167.
DOI: 10.1145/3761668.3761693