Opening the Personal Gate between Technology and Health Care
M. Giacomini et al. (Eds.)

© 2026 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI260192

Knowledge-Based Interpretation of Multi-Modal Clinical Findings: Evaluating a Local Agentic Bridge Between Worlds

Leonhard HAUPTFELD[a,1], Moritz GROB[a,b], Julia LIEPOLD[a,c], Andrea RAPPELSBERGER[b], and Klaus-Peter ADLASSNIG[a,b]
[a]*Medexter Healthcare, Borschkegasse 7/5, 1090 Vienna, Austria*
[b]*Medical University of Vienna, Center for Medical Data Science, Institute of Artificial Intelligence, Spitalgasse 23, 1090 Vienna, Austria*
[c]*TU Wien, Institute for Logic and Computation, 1040 Vienna, Austria*
ORCiD ID: Leonhard Hauptfeld 0009-0006-5902-4666, Moritz Grob 0009-0006-7867-9465, Julia Liepold 0009-0004-5416-6417, Andrea Rappelsberger 0009-0001-0345-4011, Klaus-Peter Adlassnig 0009-0008-5046-1549

Abstract. Contemporary clinical practice still produces unstructured data like free-text reports or scans, hindering automated interpretation by knowledge-based clinical decision support (CDS) systems that rely on structured data. Large language models (LLMs) show potential for interpreting such findings but face challenges in accuracy, infrastructure demands, and data privacy. Integrating LLMs with modular knowledge-based CDS systems could provide validated interpretations of such findings, but models need to call CDS modules with perfectly accurate parameters. The accuracy of multiple size classes of LLMs calling Arden Syntax Medical Logic Modules for hepatitis serology interpretation of varying complexity from unstructured multi-modal inputs is tested using a novel framework. Computationally lean LLMs like GPT-OSS were found to handle a small amount of low-complexity parameters with high accuracy, approaching clinical feasibility for private and reliable CDS interpretation of multi-modal data. Accuracy decreased sharply for tools involving more numerous or complex quantitative parameters.

Keywords. Clinical Decision Support, knowledge-based, Large Language Models, Arden Syntax, ArdenSuite, multi-modal

1. Introduction

Modern care necessitates the rapid, precise interpretation of diverse clinical findings into actionable insights. Knowledge-based CDS systems offer these insights but require structured inputs, driving extensive efforts towards data standardization. Nevertheless, clinical information still remains locked in unstructured formats such as free-text remarks or the portable document format (PDF).

The recent rise of the large language model (LLM) has resulted in a multitude of studies regarding their ability to derive and interpret information from unstructured data like documents and images, but challenges with accuracy remain [1]. Advanced models

[1] Corresponding Author: Leonhard Hauptfeld; E-mail: lh@medexter.com

require substantial computational infrastructure, raising privacy concerns if data is sent to external providers. Accuracy can be greatly improved by enabling LLMs to invoke knowledge-based interpretation tools for reasoning [2], a promising approach often referred to as agentic use [3]. Smaller, advanced models now offer this capability and enable privacy-focused hybrid reasoning within standard computing constraints.

To accurately locate interpretation tools, LLMs need to be supplied with structured metadata about each tool and its input parameters. Arden Syntax, an HL7 standard language for executable medical logic [4], features such a structured format. With minimal modifications, its medical logic modules (MLMs) can serve as drop-in interpretation tools. This integration of the structured knowledge within MLMs and the semantic understanding of computationally lean LLMs yields a novel, modular and practical bridge architecture that could eliminate the requirement for pre-structured data. This architecture facilitates improved versatility of classic CDS systems while resolving the critical infrastructure and data privacy bottlenecks inherent in clinical settings, provided its interpretation accuracy across data modalities shows levels viable for clinical use.

2. Methods

2.1. Agentic knowledge-based bridge architecture

ArdenSuite, an Arden Syntax-based CDS platform, was extended with a Model Context Protocol (MCP) server [5] to facilitate the use of MLMs as LLM tools. As a standardized interface, MCP enables LLMs to seamlessly query and execute these tools.

Each tool definition provided by MCP must contain at least a name, a description, and a specification of expected structured inputs. Already contained within each MLM is a structured definition of its name and purpose, which are mapped to tool name and description. As MLMs lack an input definition, a preliminary addition to the syntax was devised for this architecture and mapped to the tool definition. Via these mappings, every MLM installed in ArdenSuite could be made available as an LLM tool via MCP.

Upon deriving inputs from unstructured data, the LLM invokes the relevant MLM via MCP, which provides a validated textual interpretation (Fig. 1). This design isolates the LLM's role to data extraction, ensuring the MLM remains the source of clinical logic.

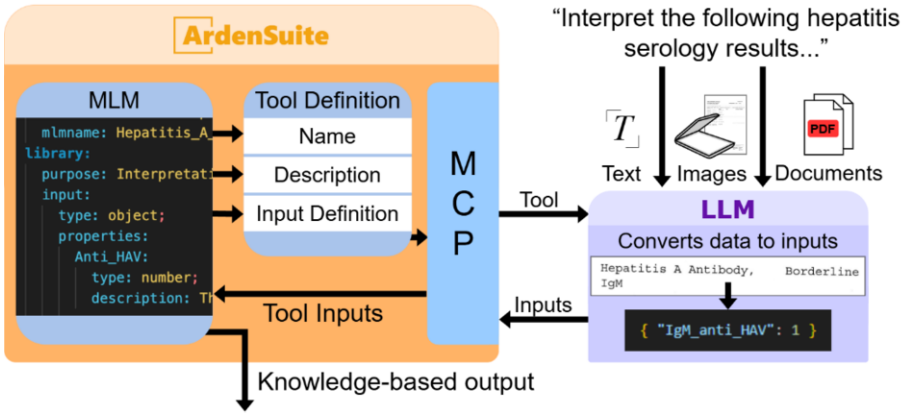

Figure 1. Interpretation process and components of the agentic knowledge system based on ArdenSuite

2.2. Hepaxpert MLMs as LLM tools

To evaluate the architecture, the MLMs of the hepatitis serology result interpretation system Hepaxpert [6] were enhanced with an input definition and loaded into ArdenSuite. Hepaxpert features three MLMs, one for each hepatitis type (A, B, and C). Each one features a varying number of qualitative and quantitative inputs, making it a good fit for evaluation. Each qualitative input is a natural number from 0 to 3 representing the test results *not tested for (0)*, *borderline (1)*, *negative (2)*, and *positive (3)*. The hepatitis A module expects three qualitative inputs, module B six, and C two. The hepatitis B module includes an additional quantitative input, the hepatitis B antibody concentration *Anti_HBs_Titre*, which is eventually resolved into the same qualitative states based on fixed concentration thresholds. Test results are provided to the LLM as qualitative descriptions (e.g., "Not tested"), which must derive them into the correct number for qualitative inputs and a value within the correct thresholds for the quantitative input.

A total of 24,656 input combinations across all types are possible, with 64 being of type A, 24,576 of type B and the remaining 16 of type C. For each combination, a synthetic text prompt and a matching faux PDF serology report were created. The number of printed and scanned image samples was capped at 100 due to time constraints and environment concerns. The image dataset included the complete type A (n=64) and type C (n=16) datasets, with 20 random type B samples.

2.3. LLMs and Evaluation Approach

As LLM tool calling accuracy generally scales with parameter size [7], three LLMs with tool calling capability of all size classes that could feasibly be hosted on premise with under 48 GB of VRAM were chosen (Table 1). Since none of these LLMs processed PDFs directly, their text was first extracted via the Kreuzberg tool before prompting. For the text-only models, the scans were converted to text using the optical character recognition engine Tesseract, while the multi-modal Qwen3-Omni processed them directly.

Table 1. Evaluated large language models.

Name	Parameters (Precision)	Modalities	Thinking
Qwen3	*1.7 billion (16-bit precision)*	*Text*	*Yes*
GPT-OSS	*20 billion (MFXP4 quantization)*	*Text*	*Yes*
Qwen3-Omni	*70 billion (AWQ8 quantization)*	*Text, Image, Audio*	*Yes (Disabled due to over-thinking)*

Inference for GPT and Qwen3 models was performed using vLLM [7] 0.11.0, while Qwen3-Omni used a maintainer-provided custom build of vLLM 0.9.3. The temperature was set at 0.6 to ensure consistent tool calls. Each LLM was evaluated for their individual accuracy with the text, PDF and image samples. Each prompt had to result in a tool call fully matching the expected tool name and parameters to count as a success. Partially correct calls were counted as failures due to them being just as irrelevant as fully incorrect ones in a clinical setting. In addition to the total result, metrics were grouped by hepatitis variant and parameter type to more precisely pinpoint any accuracy issues.

3. Results

Reviewing the results for each modality and tool (Table 2), it is evident that GPT-OSS outperformed the other models by a considerable margin, matching or exceeding their accuracy in every category. The high-parameter hepatitis B tool generally scored lower accuracy compared to the other types. Comparing its average accuracy relative to the average accuracy of hepatitis A and C across all modalities, it underperformed by 94.05% in Qwen3, by 60.38% in GPT-OSS and by 97.49% in Qwen3-Omni (Q3-Omni).

Table 2. Tool calling accuracy

Model	Text			PDF			Image		
	Hep. A	Hep. B	Hep. C	Hep. A	Hep. B	Hep. C	Hep. A	Hep. B	Hep. C
Qwen3	65.62%	14.73%	100%	20.31%	0.03%	81.25%	9.38%	0%	56.25%
GPT-OSS	100%	87.33%	100%	92.19%	11.63%	81.25%	71.88%	15.00%	93.75%
Q3-Omni	95.31%	2.16%	93.75%	81.25%	4.09%	75.00%	67.19%	0%	87.50%

When comparing the accuracy between modalities, each model shows a drop from the text prompt except for one instance with very low absolute accuracy (Table 3). Averaging the loss for all tools, Qwen3 drops 26.25 percentage points for PDF and 38.24 for image, GPT-OSS drops 34.09 for PDF and 35.57 for image and Qwen3-Omni drops 10.29 for PDF and 12.18 for image. Contrary to expectations for a model with native image processing, Qwen3-Omni did not outperform GPT-OSS, which used OCR, nor did it retain a more consistent accuracy across modalities. The accuracy drop-off is even more pronounced for the hepatitis B tool, where the relative accuracy loss consistently exceeds that of the other tools by a significant margin.

Table 3. Tool calling accuracy loss compared to text, percentage points and relative

Model	PDF			Image		
	Hep. A	Hep. B	Hep. C	Hep. A	Hep. B	Hep. C
Qwen3	-45.31 (-69.05%)	-14.70 (-99.80%)	-18.75 (-18.75%)	-56.24 (-85.71%)	-14.73 (-100%)	-43.75 (-43.75%)
GPT-OSS	-7.81 (-7.81%)	-75.70 (-86.68%)	-18.75 (-18.75%)	-28.12 (-28.12%)	-72.33 (-82.82%)	-6.25 (-6.25%)
Q3-Omni	-14.06 (-14.75%)	+1.93 (+89.35%)	-18.75 (-20.00%)	-28.12 (-29.50%)	-2.16 (-100%)	-6.25 (-6.67%)

When examining the accuracy of individual inputs, the most notable obstacle to accurate hepatitis B tool calls was the quantitative *Anti_HBs_Titre* input. With all other inputs beyond 98% accurate in the text modality for GPT-OSS, it evaluated at 91.51% accurate. In the PDF category, five qualitative inputs remained above 90% accuracy, but the quantitative input dropped to 20.28% accuracy. Relative accuracies remained similar in the image modality and across the other models.

4. Discussion

Among the evaluated models, GPT-OSS emerged as the clear choice for tool calling. Its perfect accuracy on the lower parameter hepatitis A and C tools in text mode may prove sufficiently accurate for clinical use. The smaller Qwen3 model delivered a notable performance on the low-parameter hepatitis C tool.

Accuracy was inversely correlated with tool dimensionality. Performance degraded as input counts increased (type C to A) and when integrating a higher complexity quantitative variable (type B), although the generally high input count of type B should be considered as a covariate. This appears to be a much more limiting factor to tool calling accuracy than a language model's ability to understand multi-modal data, although their drop in accuracy from a plain text prompt is still notable.

For the only truly natively multi-modal model, Qwen3-Omni, higher accuracy on image samples was expected, but it performed worse on them than GPT-OSS. Further testing is needed to see if this holds true for other multi-modal models or is specific to Qwen3-Omni. Disabling its reasoning mode and quantization may have affected the results but was necessary to keep within the computing resource constraints.

5. Conclusions

The devised knowledge-based agentic bridge architecture showed promise working with unstructured text, but still struggled in the more interesting document and image modalities. In prior research, fine-tuning has proven a viable way of increasing tool calling accuracy with complex tools [8]. This fine-tuning could be automated by inferring possible MLM input combinations and yield validated LLMs for certain knowledge-bases. As tools with few inputs yield the most promising accuracy results for clinical use, they should be the focus of ongoing research efforts, especially in a clinical setting with authentic data. This architecture also dedicates the LLMs to the task of structuring unstructured data, and relies on MLMs for textual output. Feeding the output of MLMs back into LLMs to potentially improve their accuracy and generate richer textual interpretations is worthy of further study. Despite constraints on scenario complexity and modality imposed by contemporary LLMs, these findings demonstrate the system's future viability in knowledge-based CDS with clinical-grade accuracy.

References

[1] L. Buess, M. Keicher, N. Navab, A. Maier, and S. Tayebi Arasteh, From large language models to multimodal AI: a scoping review on the potential of generative AI in medicine, *Biomed. Eng. Lett.* **15** (2025) 845–863. doi:10.1007/s13534-025-00497-1.

[2] A.J. Goodell, S.N. Chu, D. Rouholiman, and L.F. Chu, Large language model agents can use tools to perform clinical calculations, *Npj Digit. Med.* **8** (2025) 163. doi:10.1038/s41746-025-01475-8.

[3] X. Xu, and R. Sankar, Large Language Model Agents for Biomedicine: A Comprehensive Review of Methods, Evaluations, Challenges, and Future Directions, *Information*. **16** (2025) 894. doi:10.3390/info16100894.

[4] HL7 International, Health Level Seven Arden Syntax for Medical Logic Systems: Edition 3.0, STU2, (2025). https://www.hl7.org/implement/standards/product_brief.cfm?product_id=639 (accessed April 25, 2025).

[5] Anthropic, Model Context Protocol, *Model Context Protocol*. (n.d.). https://modelcontextprotocol.io/ (accessed November 1, 2025).

[6] K.P. Adlassnig, C. Chizzali-Bonfadin, M. Kreihsl, F. Schulz, W. Horak, and H. Hofmann, HEPAXPERT-III: knowledge-based interpretation of serologic tests for hepatitis A, B, C, and D, *Medinfo*. **8 Pt 2** (1995) 1683.

[7] S. Badshah, and H. Sajjad, Quantifying the Capabilities of LLMs across Scale and Precision, (2024). doi:10.48550/arXiv.2405.03146.

[8] G.A. Manduzio, F.A. Galatolo, M.G.C.A. Cimino, E.P. Scilingo, and L. Cominelli, Improving Small-Scale Large Language Models Function Calling for Reasoning Tasks, (2024). doi:10.48550/arXiv.2410.18890.